PRETHESIS REPORT

DEVELOPMENT OF AN ECG SCORING SYSTEM APPLIED FOR DETECTION AND QUANTIFICATION OF MYOCARDIAL DAMAGE USING MORPHOLOGICAL FEATURES AND DE-TRENDED FLUCTUATION ANALYSIS

INSTRUCTOR:
Dr. Le Quoc Trung

STUDENT:
Pham Khoi Nguyen

Project name

Development of an EKG scoring system applied for detection and quantification of myocardial damage using morphological features and de-trended fluctuation analysis

Contact Information

Name: Phạm Khôi Nguyên

School: International University – Vietnam National University (VNU)

Department: Biomedical Engineering Year: 2013 - 2017

Email: phamkhoinguyen1995@gmail.com Mobile Phone: + (84) 914 118 896

Mailing Address: Room 34, Binh Phu Apartment, Ward 10, District 6, Ho Chi Minh City

Project Adviser

Name: Dr. Lê Quốc Trung

Title: PhD, University Lecturer Organization: International University – VNU

Email: lequoctrung@gmail.com Mobile phone: + (84) 915 538 938

Development of an EKG scoring system applied for detection and quantification of myocardial damage using morphological features and de-trended fluctuation analysis

Abstract: During the last decade we have witnessed the tremendous growth of Telecommunication technologies, particularly the advancement in the field of semiconductors, wireless network and cloud infrastructure that could eventually bring forward the era of homecare [3-5]. As a result, these innovations have made early diagnosis and prevention medicine for Cardiovascular Disease (*CD*) feasible, something that could not be happened 20 years ago [4]. In Vietnam, there is a constant interest in the development of various types of real - time, wireless EKG (ECG) devices dedicated to people who are at high risk of CD, where many startups emerged to provide this type of innovative products to the community [12, 13]. As appealing as it may sound, however, these devices are currently lack of a decent diagnosing system. They are only capable of measuring and displaying data, some managed to transmit data to clinical centers via the Internet and the diagnosis is still carried out by doctors [12]. Therefore, this project aims to develop an algorithm capable of analyzing EKG input to provide automatic diagnosis of Cardiovascular Disease. In future work, the algorithm will be integrated into an online platform dedicated for real – time diagnosis of CD. In this paper, literature review on Cardiovascular Disease, conventional and novel diagnosis methods, development of the algorithm (a risk score system) and result validation will be covered. Implementation of algorithm into an online system will be described in the final thesis. In conclusion, the proposed algorithm strives to provide doctors with a remote diagnosis tool for Cardiovascular Diseases, which will help ease the overloaded condition at the hospital, save cost for treatment and bring peace of mind for people who are at risk of CD in Vietnam.

Table of contents

LITERATURE REVIEW

1.1 CURRENT CONDITION

For many decades, Cardiovascular Disease (CD) is the leading cause of hospital mortality in many countries around the world, including Vietnam [8]. As reported by the World Health Organization (WHO), there are about 17.3 million people died from Cardiovascular Disease in 2008, representing about 30% of all global cases of death. Not only that, this value is expected to rise to 23 million in the early 2030s [8]. The World Heart Federation (WHF) estimates that the incidence of Cardiovascular Disease in Vietnam in 2017 could reach 20% of the total adult and elder population, ranking the fourth highest in the world.

Cardiovascular Disease poses a direct threat to the health and lives of many people, among them the middle-aged group and elderly people have the highest vulnerability [10]. In Vietnam, this problem becomes more serious as the majority of the elderly and middle-aged people also suffer from high blood pressure, which is the number one cause of Cardiovascular Disease and also the cause of death of more than 7 million people each year. According to the latest survey of Vietnam Cardiology Department in 2016 [21], approximately 48% of all adults will develop hypertension (information extracted from Hypertension Conference Second Vietnam 2016 in Hanoi with the theme "multidisciplinary approach to hypertension"). The statistics below also demonstrates the worsen condition of Hypertension and Cardiovascular Disease in Vietnam:

- 1960: 1% of the total number of middle-aged people in northern Vietnam.
- 1976: 1.9% of middle-aged people in northern Vietnam.
- 1992: 11.7% of middle-aged people in the entire country.
- 2001: accounting for 23.06% of the total patients in Hanoi particularly.
- 2007: accounting for 16.32% of the total patients across the country.

Therefore, finding solutions to prevent the "silent killers", namely high blood pressure, heart attack and stroke, has long become one of the most urgent issues in national healthcare.

According to many scientists, the incidence of developing these diseases is higher in urban areas than in rural areas [8]. Explaining this phenomenon, Prof. Pham Gia Khai, Chairman of the National Heart Association of Vietnam, said that the unhealthy lifestyle (lack of physical exercise) and eating habit is the main cause for this condition.

Moreover, hospital system in the city is under a lot of pressure because of the steady increase in the number of patients every year [8, 10]. As the Ministry of Health stated, the ratio between the amount of doctors and pharmacists over the amount of patients is currently 7.61 and 2.2 out of 1,000. Through this value we could see that besides the progression of Cardiovascular Disease, the hospital overloaded condition is also another major social challenge that need decent solutions.

With this in mind, the aim of this project is to develop a platform capable of performing automatic diagnosis of Heart Disease and thus allowing and helping doctors to provide homecare solution to patients who are at high risk of CD in Vietnam.

1.2 MYOCARDIAL INFARCTION

Acute Myocardial Infarction (AMI), also known as Heart Attack, is a disease caused by insufficiency of blood supply to the heart's tissue [33 - 39]. Generally, heart's tissue is supported by a system of blood vessels. When these blood vessels suffer from Coronary Artery Disease (CAD) – the constriction of the artery that obstruct blood flow by the formation of fat and cholesterol beneath the vessel's inner wall, some part of the heart does not receive enough blood supply. This phenomenon, if left untreated for a period of time, can eventually lead to cells death. The condition when some region of the heart died because of the derivation of blood supply and cannot function normally is called Myocardial Infarction.

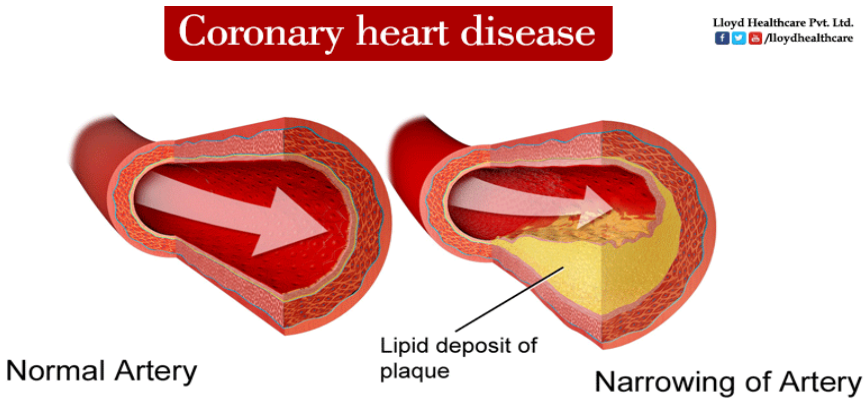


Figure 1: *The most common cause of cardiovascular disease is the formation of a lipid deposit that obstructs artery blood flow, causing an insufficiency of blood supply to the myocardium that eventually leads to cardiovascular damage.*

Acute Myocardial Infarction is the leading cause of death during patients' hospitalization in the United State. It has been reported that each year in the America, 1.1 millions of people suffer from Myocardial Infarction and haft of them get an acute attack [4]. In Vietnam, the number of patient who suffered from AMI tends to increase drastically during the last 20 years (from 1980 to 2000): from 1980 to 1990 there was 108 patients, from 1991 to 1995 there was 92 patients and from 2000 to 2001 the number escalated quickly: more than 1.500 patients [4] and 17.4% (261 patients) died [4]. Most of these patients are elderly (>65 years old) whose biological characteristics of coronary artery make them more susceptible to AMI.

1.3 CONVENTIONAL DIAGNOSIS METHODS

Treatments for Acute Myocardial Infarction (AMI): Stenting and Coronary Bypass Graft for example, are extremely time – consuming and expensive [6]. Therefore, the current diagnosis techniques focus on early detection of AMI before tissue death occurs. Many powerful medical diagnosis techniques, including Cardiac MRI [4], CT Angiography [5] and Echocardiography [1], have been used tremendously for the diagnosis of AMI. These diagnosis techniques have been known to be extremely accurate in addition to providing high resolution medical images about the inner structure and function of the myocardium.

1.3.1 CARDIAC MRI

Cardiac MRI utilizes Magnetic Resonant characteristics of the Hydrogen atom within the myocardium to create images [18, 19]. Generally, hydrogen atom naturally exhibits a chaotic orbiting behavior when it fluctuates and rotates in various direction around its imaginary central axis. When being subjected under strong magnetic field, these hydrogen atoms are lined up in a direction that is either parallel or anti – parallel to the direction of the applied magnetic field [18]. When the magnetic field is suddenly stopped, these atoms instantaneously pound back to their normal state, emitting magnetic wave during the process that is eventually collected by a detector to form the image. Since hydrogen atom is the primary component of soft tissues such as the heart, Cardiac MRI has been vastly used as diagnosis technique for cardiovascular disease in the recent years [19].

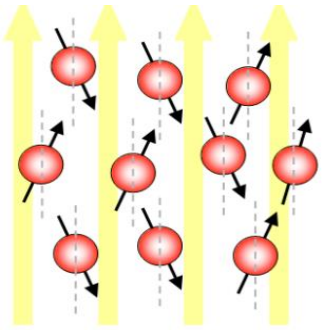


***Figure 2**: each hydrogen atom exhibits a rotation around itself while revolving around an imaginary axis that is either parallel or anti – parallel to the direction of the applied magnetic field.*

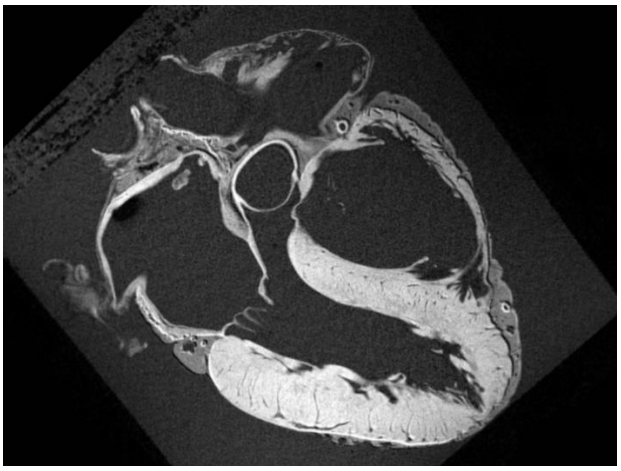


***Figure 3**: after the atoms pound back to their normal state, the emitted magnetic radiation is captured to form the image. MRI image comes with extremely high resolution.*

The images captured from these devices come with extremely high resolution. Cardiac MRI can clearly demonstrate the underlying structure of the heart as shown in *figure 3*. Therefore, the Cardiac MRI images are normally used for detection of the damaged area [18]. It is in fact the most advanced diagnosis technique for Cardiovascular Disease nowadays. Cardiac MRI does, however, suffer from many major drawbacks. It is one of the most extremely expensive diagnosis techniques and it also requires physician with immense clinical practice to perform the test [19]. Clinical preparation is also very time - consuming and the device is not always available in many hospitals.

1.3.2 CT ANGIOGRAPHY

CT Angiography utilizes X-rays absorption properties of the artery tissues to illuminate coronary pathway [21]. When performing the test, a contrast agent dedicated to improve image quality is injected to the vein. Then, blood flow carries these particles back to the heart coronary system, where they absorb X-rays beam generated by the instrument. The intensity of the beams transmitted through the artery is captured by a detector and the image of the coronary path way is formed. CT angiography is a strong diagnosis technique when it comes to the detection of obstructed coronary artery as shown in *figure 4*.

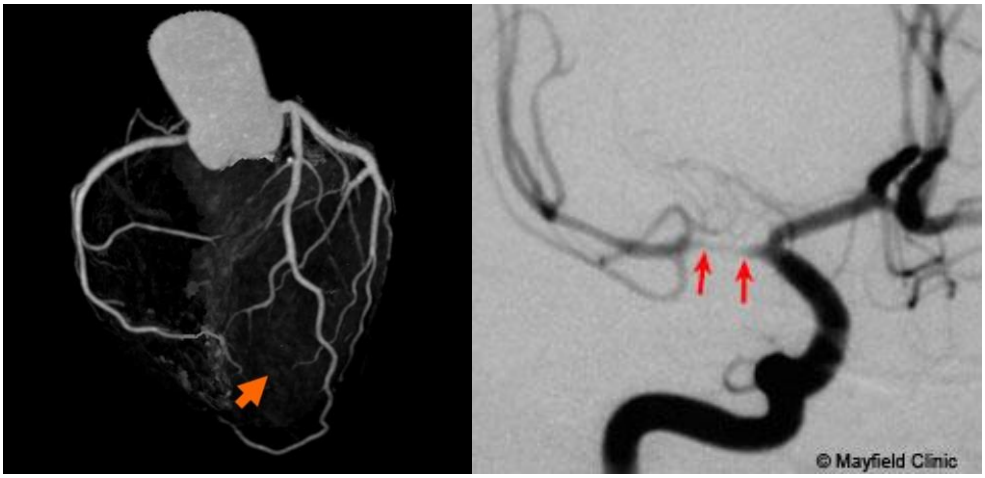


***Figure 4**: coronary obstruction detected by CT angiography*

CT angiography, however, also suffers from many disadvantages. First, this diagnosis is only performed when CD is highly suspected because the subject is at risk of radiation exposure. The injected contrast agent also causes allergy in many patients and the test also comes with high cost.

1.1.1 ECHOCARDIOGRAPHY

Echocardiography is the technique that utilizes ultrasonic waves to visualize the surface area and inner structure of the heart [26]. When performing the diagnosis, an ultrasonic probe is directed through the mount down into the throat to reach the region near the heart. Then, ultrasonic beam is generated. The sound waves travel to the myocardium, some got transmitted and some got reflected. The reflected wound wave is then collected by the probe and through piezo electricity technique, ultrasonic wave is converted into electrical current for diagnosis and imaging. The image captured could visualize the inner structure and also capable of demonstrating the pumping ability of the heart. Echocardiography is majorly used to inspected the damaged areas by looking at the region that fails to contract during each heartbeat [26 - 27].

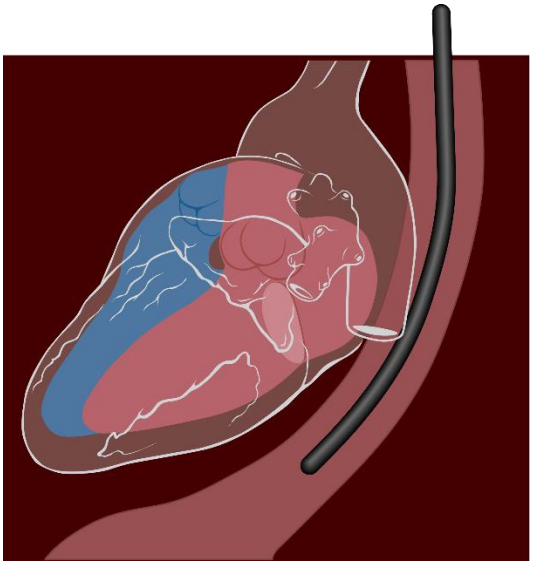


***Figure 5**: the ultrasonic probe is directed from the throat to reach the region as close as possilble to the heart*

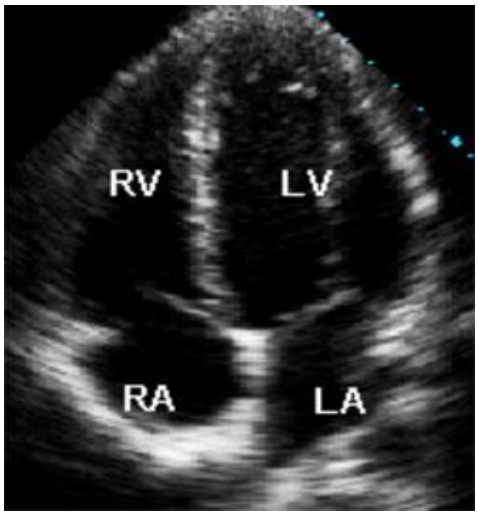


***Figure 6**: image of the inner structure of the heart captured during each heartbeat*

Among these techniques, echocardiography is the cheapest and simplest diagnosis technique of CD. Its main drawbacks are the discomfort during the test and the image resolution is not as high as the other two.

1.3.3 SUMMARY AND COMPARISION

In summary, the invasive medical diagnosis techniques of CD have both advantages and disadvantages. Each type of these test strives to diagnose different characteristics of the heart such as CD caused by dead region of the myocardium (Cardiac MRI), by the obstructed coronary system (CT angiography) or by failure of the pumping ability (Echocardiography). The table below provides a short summary and comparison of these diagnosis techniques.

	CT Angiography	Echocardiography	MRI
Dependencies	ECG gated to obtain optimal image resolution	ECG gated	Is not dependent on ECG but usually taken with ECG
Time requirement	2 – 3 hours including preparation	20 – 30m + 10 – 20m (with Doppler effect)	30m, or 50 – 60m for contrast enhancement
Additional drugs	❑ Beta block ❑ Nitroglycerin ❑ Contrast agent ❑ Safety medication	❑ White gel	❑ Contrast enhancement agent
Complication	❑ Allergy ❑ Radiation exposure	❑ None	❑ None
ED Standard Procedure	Recently being proposed	Used for advance diagnosis	Re – test for healed MI
Effectiveness	Specific location can be point out	Specific location can be pointed out with echocardiography	Specific, 3D model of heart
	Information are less likely to be missed [7][10]	Additional info: ❑ Size and shape ❑ Function ❑ Tissue damage	Additional info: ❑ Size and shape ❑ Function ❑ Tissue damage

Usage	❑ Infarct location ❑ Coronary blockage	❑ Infarct location ❑ Coronary blockage (low sensitivity) ❑ Heart activity ❑ Tissue damage ❑ Thrombolysis	❑ Infarct location ❑ Coronary blockage ❑ Heart activity ❑ Reversible vs irreversible MI
Cost	High cost but take only one measurement	High cost due to physician with immense medical practice	High cost

1.4 ELECTROCARDIOGRAM SIGNAL

ECG, or Electrocardiogram, measures the electrical activity of the heart during each consecutive heart beats. The current clinical diagnostic technique using ECG analyses the shape of the waveform and calculate the magnitude, energy and entropy of the signal to deliver valuable information about the heart. For example, by focusing on some specific segments of the signal: P wave, T wave, the presence of Q wave and ST segment, detection of myocardial infarction, cardiac arrest and arrhythmia can be achieved.

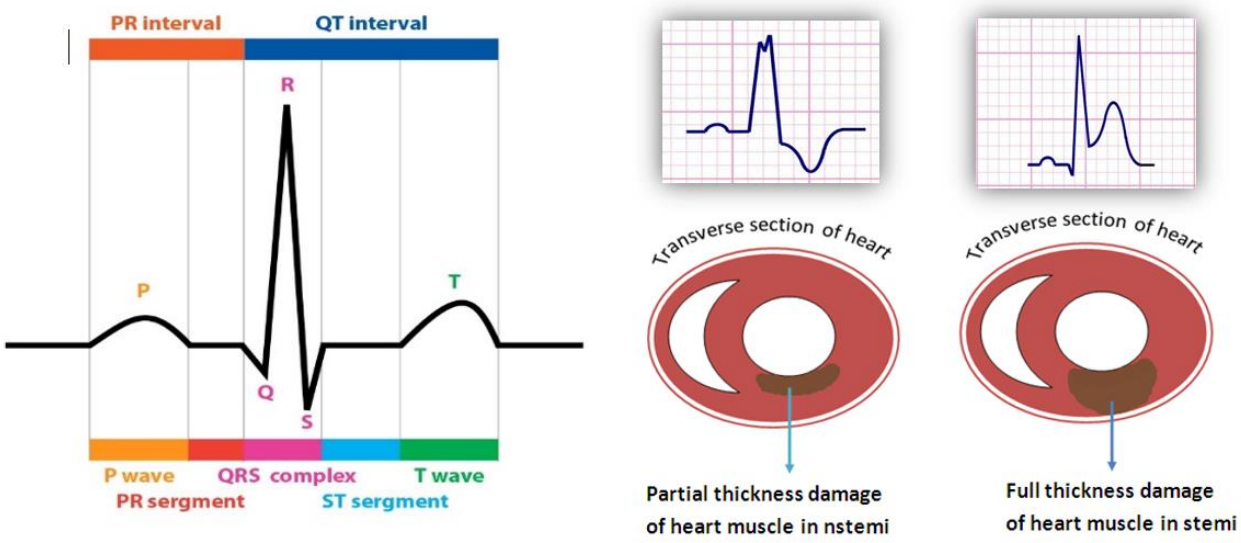


***Figure 7**: from left to right are a normal ECG waveform versus two altered ECG waveforms that correspond to different types of cardiovascular problem*

ECG measurement has been used as a standard procedure for approving patients to the Heart Disease Department in almost hospitals around the world because of its low expense, fast and reliable. In addition to that, technical aspects of the signal also elevate ECG as a wonderful approach to develop a small, light-weight device that is suitable for home care solution. To illustrate this point, many famous chipset companies such as Texas Instrument and National Instrument, is currently providing small, affordable ECG modules with very good signal quality. Most importantly, the greatest interesting feature of this signal remains in its medical prognostic value. The use of ECG to forecast the occurrence of heart attack is still an uncultured field but yet extremely profitable if it was

discovered. The model can help elder people prevent the occurrence of AMI or make immediate responses to sudden cardiac attack.

1.4.1 SIGNAL PHYSIOLOGY

Figure 7 above represents a typical example of the electrical activity of the heart. ECG (Electrocardiogram) is a sequence of wave forms that manifests the dynamical activity of the heart during consecutive heart beats. Each part of the wave form corresponds to a specific function in a specific region of the heart.

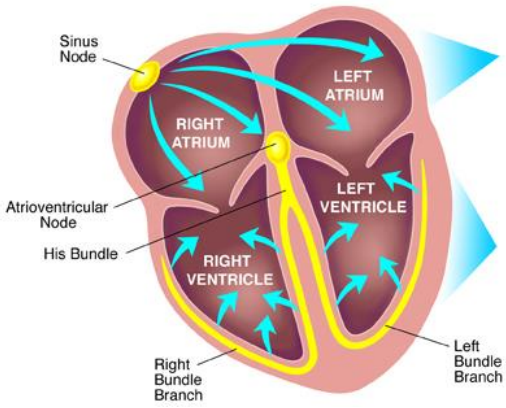


***Figure 8**: Electrical conduction system of the heart*

Our heart consists of 4 champers: 2 atrial receiving blood from the body and 2 ventricles that receive blood from the atrial and pump back to the body. The pumping activity of these champers is described as the peak value in the ECG wave form, where the first peak (P wave) represents for the depolarization and contraction of the atrial, and the second peak (R wave) represents for the depolarization and contraction of the ventricle. The final peak (T wave) manifests the repolarization, or relaxation, of the ventricle thus preparing the heart for the next heart beats.

When analyzing ECG signal, the combination interpretation of waves together with the interval between these waves can yield reliable information about the overall activity of the heart. Each normal heart beat consists of a P wave, a QRS complex and a T wave and the corresponding interval in between. In figure 1.2, the electrical current that flows through the heart in each consecutive heart beat is described, providing the physiology perspective of the intervals.

First an action potential is generated at the SA node, producing an electrical current that flows to the atrial. This electrical current stimulates the contraction of the atrial, resulting in the formation of P wave in the ECG signal. The current then flow through the atrial to the AV node, where it then travels to the left and right branch of the ventricle by conductive fibers known as left bundle branch and right bundle branch. The arrival current then stimulates the contraction on both the left and right ventricle, representing the R peaks in QRS complex. It is also important to mention that during this interval, the simultaneous contraction of the ventricle and relaxation of the atrial contribute to the complexity of the QRS pattern. After this point, ST segment represents for the time delay between ventricle contraction and relaxation, where all the ventricular tissues, after contraction, tend to bounce back to normal in order to be ready for the repolarization. This is the most important interval for spotting any damage to the ventricle tissues because the electrical current developed by any injury or inflammation to the ventricle appears clearly during this isoelectric process. Finally, the T wave occurs as the result of ventricle repolarization, or relaxation.

1.4.2 CHARACTERISTIC WAVES

As previously mentioned, the ST segment is the most valuable part of the ECG waveform to identify any damage to the ventricle, where most AMI occur. Beside ST segment, other valuable waves that also need attention is the T wave and Q wave. Acute Myocardial Infarction manifestation in ECG signal is a sequence of changes in the T wave, ST segment and Q wave as described in figure 7. Firstly, the constricted blood vessels prevent or decreases blood flow to some specific region of the heart. Atrial tend to have larger coronary vessels, therefore it is less susceptible to blood derivation. However, the ventricle is supported by a complicated vessel system, some is really big (Bundle of His) and some is really small (Purkinje Fibers), therefore the restriction of blood supply is more viable, thus making the ventricle more susceptible to blood derivation. When the tissue is lack of blood supply, the Ischemic Event occurs, resulting in the injury or inflammation in some region of the heart. These injury generates addition electrical current that can be detected during isoelectric process of the heart: the ST segment. During this phase, T wave first becomes peaked and then ST changes occur. If the ST segment elevates, the myocardium is interpreted as having full thickness damage of the heart muscle (Figure 7b). If the ST segment depresses, then the traverse damage can be the cause (Figure 7c).

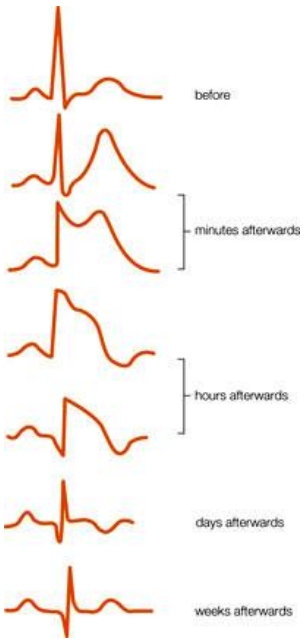


***Figure 9**: Dynamical changes of ECG waveform during the formation of AMI*

Finally, if the disease is left untreated for a long period of time, tissues death will eventually occur. The formation of a pathological Q wave (larger and more negative Q wave) also develops during this stage. This is in fact the final stage of Acute Myocardial Infarction.

In conclusion, if the ST segment are elevated representing tissue injury, the phenomenon is categorized as ST Segment Elevation Acute Myocardial Infarction (STEMI). If the ST are horizontally normal or depressed representing ischemic event or tissue injury, it is categorized as Non – ST Segment Elevation Myocardial Infarction (NSTEMI).

1.5 ANALYSIS TECHNIQUES APPLIED IN MEDICAL AND CLINICAL APPLICATIONS

Beside the medical point of view, the advancement in Information Technology has introduced the birth of Big Data. Despite the availability of tremendous amount of data nowadays, very little useful information has been derived that can be turned into practical knowledge. Within this condition, Data Mining plays an essential role in how we make use of these databases. Data Mining is the process in which useful information can be analyzed and turned into practical knowledge.

Regarding Medical Application, medical data can be analyzed to produces significant information that can be used for Diagnosis and Predictive purposes. In our research interest, we strive to integrate Data Mining into the research methodology in order to study and understand the hidden dynamical process that lies within the ECG signal itself. To be more elaborate, large amount of ECG signals from various databases will be taken into account, where different analysis techniques are applied to find the underlying features that eventually lead to Acute Myocardial Infarction. In this paper, these features are referred as the Hidden Prognostic Value of ECG.

While ECG is observed quite as a periodic process since the consecutive waveforms tend to repeat over the time, the overall signal is still considered as a chaotic, nondeterministic system. In order to adequately analyze ECG signal using probability and statistic technique, several requirements have to be taken into account:

1. The technique must be able to express the randomness of the signal.
2. While the data is chaotic and nondeterministic, periodicity must be observable.
3. Transition Probability of one stage to another has to correspond with physiological meanings.

During the analysis process to withdrawn the underlying ECG features, it is important to turn ECG signal from Time Series Domain into different other domains such as: Frequency Domain and Phase Space Domain.

1.5.1 ANALYSIS DOMAINS

a. Time Series domain

In Time Series Domain, ECG signal is a sequence of measurement over the time. This domain gives the most basic information about the electrical activity of the heart through different cardiac circle and information regarding these activities is manifested in the shape of the waveform as previously described in section 1.4. In order words, this domain allow us to withdraw information about the physiological activity of the heart over the time by interpreting the shape of the waveform. However, this technique depends heavily in the visual ability to spot out the unusual patterns of the signal, not taken into account the dynamical changes that caused the formation of these abnormalities. Although applying this visual technique in detecting the abnormalities within these signal yields appropriate information for detection of AMI, it is still not an appropriate approach for prediction because at this point, clinical symptoms have already occurred.

b. Frequency Domain

The process of transforming ECG signal from time series domain into frequency domain is known as the Fourier Transform where the equation is described as following:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx,$$

In this equation, *f(x)* is the modelling function in Time Series Domain. The transformation allows the breakdown of the original signal into its frequency components versus the signal amplitude.

In reality, the modelling function of the signal is not available or too complicated to calculate. Then Fast Fourier Transform (FFT) is applied directly to a Time Series data set and turns it into frequency components. In this domain, information derived is the frequency of specific sinusoid components within the original signal and usually this technique is applied to withdraw the desired signal from noisy input. After the range of frequency of interest has been selected and further processed, the inversion technique is applied to return the signal into the original Time Series domain:

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{2\pi i \xi x}\, d\xi,$$

After the reversed transformation, visual technique is the applied to spot out the unusual waveform that corresponds with AMI.

c. Phase Space Domain

Phase space domain represents for every possible states in which the signal can be found therein. In this domain, each factors that contribute to the formation of the signal is categorized as a parameter, and each parameter is described as an imaginary axis in the phase space diagram. The state of the system is presented as a unique point within that diagram.

The transformation of a signal into state space domain is particularly useful for analyzing the periodicity of the signal. Therefore, in our research methodology, we perform state space analysis in order to understand what truly contributes to the dynamical process of the heart that make the signal repeats over the time. However, in order to perform phase space transformation, it is crucial that the state equation of the system must be known. This is not usually the case because as previously discussed, the model that represent ECG signal is nondeterministic. Therefore, the state equation is not available or too complicated to compute. In this case probabilistic technique will be applied and it works in a similar manner to phase space method. Relevant studies will be discussed in the next section.

1.5.2 ANALYSIS TECHNIQUES

This section covers some famous analysis techniques that have been tremendously cultured with ECG. The main purposes of this section is to provide a broad view about many possibilities of applied statistics and mathematics in the field of medical application and diagnostic. These techniques will also be applied in this research. The implementation and result can be found in section II and III respectively.

a. Logistic Regression

Logistic regression is a parametric model that calculates the probability for an event to happen basing on the data that has been captured in the past known as experience [34]. By applying logistic calculation, this model can analyze the strength of the relationship between categorized dependent parameters and the desired variable. After performing this calculation, Logistic Regression model can determine the core parameters that greatly affect the output of the system, and in addition, provide

a Logistic Function to calculate the output using these core parameters. When applied in clinical studies, Logistic Regression proves to be very powerful in detection of cardiac disease and predicting mortality after hospital discharge basing on some crucial information regarding patients' health [10].

One remarkable application of Logistic Regression application in the field of medical diagnosis is the development of a simple risk score for assessing clinical severity of Acute Myocardial Infarction after Hospitalization [35], by Jacob, PhD and Henry, MD. This article strives to evaluate long term mortality risk for patient with acute myocardial infarction after hospital discharge within 6 years. The study found out strong correlation between mortality and clinical parameters including shock, heart failure, ECG finding, kidney function, and age. In general, patients who have their risk score greater than 16 points are 22 times more likely to die within the next 6 years than whose score ranges from 0 to 1. The result was compared with actual death certificate and the model proved to be very accurate.

b. Artificial Neural Network

Artificial Neural Network (ANN) is the general term for a group of Biological Neural Networks models that mimics human cognitive ability to detect an event or to make future predictions basing on the past experience [36]. Similar to Logistic Regression, ANN models also learn how to calculate the provided inputs in order to give the final estimate. However, the main difference is that while Logistic Regression uses Logistic Calculation to analyze the strength of parameters' relationship, ANN treats each input parameters as an interconnected neuron that exchanges information with one another as shown in the figure below.

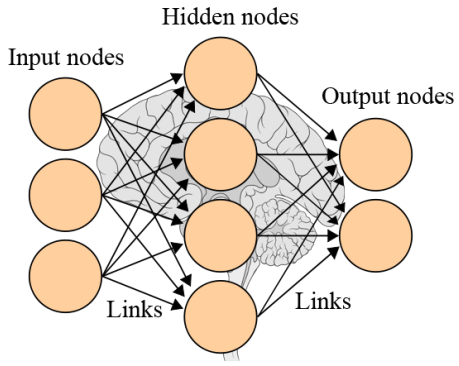


Figure 10: *The Artificial Neural Network model*

Each neuron also contains an adaptive weight representing for its degree of importance on the final result [37]. Then by applying a predefined Training Function, ANN can calculate the desired parameter basing on all of the input parameters. This ability, therefore, allow ANN to base the calibration on very large amount of available input parameters which render other forecasting models poorly performed because of the tremendous calibration associated [38]. However, its strength also implies its weakness. In order to perform ANN with tremendous amount of input parameters, a computer system with strong processing power and large data storage must be used [38], thus making these models not available for small scale analysis.

In real - life application, ANN excels as a method for classification [34], pattern recognition [39], data processing and robotic control [40]. In the field of time series analysis for clinical diagnosis and prognostic, ANN thrives as a long term forecasting models that predict accurately the outcome of diseases given with large amount of input parameters representing patient current condition. One noticeable study is the "Prediction of protein stability changes upon single-point mutations" [41], as described by Emidio, Pierro and Rita. This study involves creating an ANN system on top of a dataset of 1,615 mutations documented with numerous input parameters and outcomes. In final, this model was capable of analyzing the whole system and giving a prediction up to 90% in accuracy about the changes of protein stability. In another study, the author Stephan and Lucia make "A comparison about the methodology and clinical application of Logistic Regression and Artificial Neural Network" [34]. The result is quite interesting, where the final conclusion is that ANN is the generalized version of Logistic Regression and both perform well in the field of Biomedical Diagnosis. However, one worth mentioning weakness of ANN over Logistic Regression is that, as described above, the former takes up much more computer resources for calibration than the latter.

c. k-nearest neighbor approximation

For a robust description, k-NN approximation method is a nonparametric models used for classification and regression. The method is powerful yet simple to apply. The principle of the technique is that similar inputs will create similar outputs and is utilized to provide estimate for time

series. When applied in clinical situation for long term forecasting, k-NN analyzes the trend of the past data and create a collection of dataset categorized with some similar properties [42]. This model then determines the most similar data point to the current data point in order to perform detection or provides the next stage of this data point as the prediction. The advantage of this model is that k-NN approximation model is easy to apply and the associated hardware system does not need strong computational power. However, the drawback is that it is not complex and dynamic enough for many chaotic and nondeterministic systems such as biological systems.

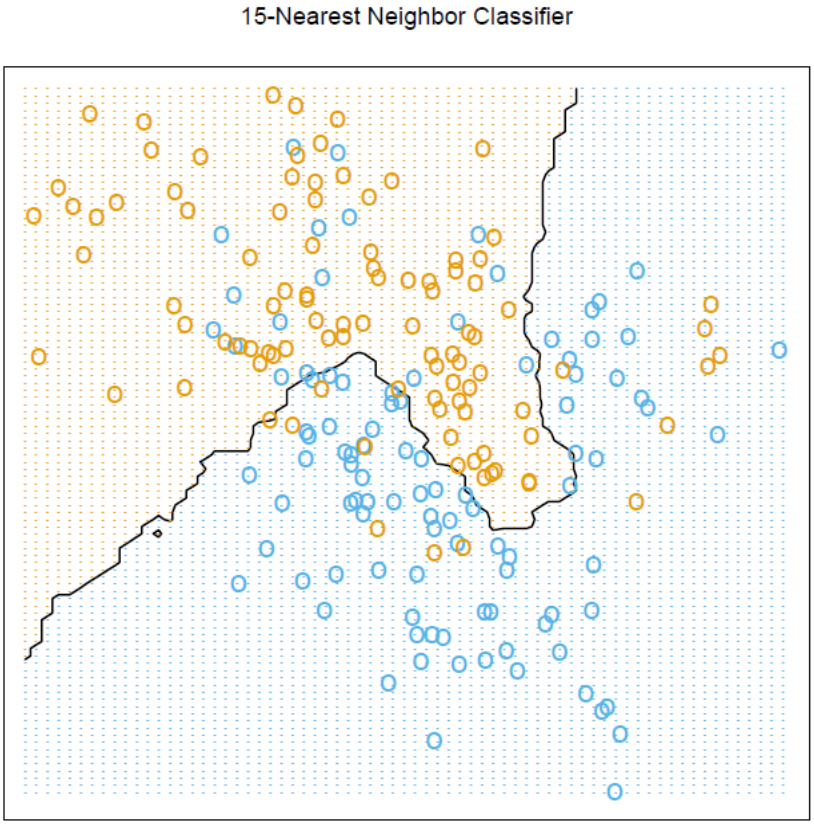


***Figure 11**: An example of k-NN analysis with data points either belong to the red class or blue class. When new data point is computed, its 15 nearest neighbors are looked up in order to decide which class this new point belongs to.*

d. OTHER RELEVENT STUDIES AND ANALYSIS TECHNQUES

The following table describes briefly many relevant studies that utilizes ECG signal to perform detection of various types of cardiac diseases. The main purpose of this section is to provide even a wider range of application that fuels the trend of using ECG as the primary source of signal for diagnosis of CD. Some of these techniques also have inspiration on the scope of this research.

Tech	Features	Description	Disease	Accuracy	Note
Spectral turbulence analysis of SAECG	Mean peaks per slice (MPPS); low-segment correlation ratio (LSCR); intersegment correlation mean (ISCM); intersegment correlation standard deviation (ISCSD); and spectral entropy (SE). The spectral turbulence analysis was considered abnormal when at least 3 of the 4 indices were abnormal: LSCR > 73; ISCM < 92; ISCSD > 105; and SE > 14.	Averaged X-Y-Z lead on the segment starting 25ms before the QRS onset and ending 125ms after the QRS offset. This segment was divided into overlapping 24ms slices in 2ms steps. Each time slice was multiplied by a 4-pole Blackman-Harris window and analyzed using the fast Fourier transformation. In order to detect the abrupt changes in activation wave-front velocity caused by abnormal myocardial regions Computing the positive predictive characteristics (PPCs), that is, curves expressing the dependence of positive predictive accuracy (i.e., the ratio [true positive]/[true positive + false positive]) on sensitivity for these 5 features	ischemic ventricular tachycardia, arrhythmic events, sudden arrhythmic death, cardiac death	Optimal criteria for risk stratification after myocardial infarction. These criteria are as follow: MPPS > 36; LSCR > 68; ISCM < 90; ISCSD > 136; and SE > 13, with the strategy requiring at least three indices to be positive for a positive diagnosis	orthogonal X,Y,Z leads using a Model 1200 EPX Arrhythmia Research Technology (Austin, TX, USA) recorder
Time series analysis of SAECG	Three conventional time domain indices were calculated: the duration of the total QRS complex (tQRS); the duration of the terminal low-amplitude signals < 40*fj, Y* (LAS40); and the root mean square voltage of the last 40ms of the QRS complex (RMS40) Considered abnormal when at least 2 of 3 variables were out of rangers: tQRS > 114ms; LAS40 > 38ms and RMS40 < 20*fiW*	Computing the positive predictive characteristics (PPCs), that is, curves expressing the dependence of positive predictive accuracy (i.e., the ratio [true positive]/[true positive + false positive]) on sensitivity for these 3 features	ischemic ventricular tachycardia, arrhythmic events, sudden arrhythmic death, cardiac death	Spectral turbulence analysis of the SAECG was a better predictor of cardiac death than time-domain analysis. However, the two methods were equivalent for the prediction of ventricular tachycardia, sudden arrhythmic death, and arrhythmic events	

Tech	Features	Description	Disease	Accuracy	Note
Logistic regression of SAECG, Holter, Radionuclide Ventriculography	SAECG: magnitude of voltage signal in the last 40ms of the filtered QRS, duration of QRS Holter: Lown Grade system Radionuclide Ventriculography used to assess ventricular ejection fraction	SAECG: a low voltage signal in the last 40ms (<40uV) of the filtered QRS complex, a long filtered QRS complex (>120ms) Lown Grade of Holter: Complex ventricular ectopic (3-5), frequent ventricular (>10), Non-sustained ventricular tachycardia (>3 + fast HR: 120/min) Ventriculography: ventricular ejection fraction <40% More information, see **table 2, figure 3.1**	ventricular tachycardia, left ventricular dysfunction, complex ventricular ectopic activity	An equation is generated that allows assessment of risk : $P(AE) = \frac{e^{\text{logit } y}}{1 + e^{\text{logit } y}}$ The finding of an abnormal SAECG in the presence of an ejection fraction <40% identified patients with a 34% probability of arrhythmic events, associated with a sensitivity of 80% and a specificity of 89%	Data analysis was performed using Student's t test, and the chi- square method 210 patients
Correlation-analysis of the clustered ECG waveforms	QRS detection algorithm, RR intervals clustering technique, T-wave and P-wave detection algorithm	The compete detection of T-wave and Q-wave: 1. QRS detection algorithm (noise robust) to create RR intervals 2. Clusters of RR intervals are created with the time-requirement (t < threshold) and geometry-requirement (mean-deviation < threshold, deviation of the deviation-curve < threshold, amplitude and duration of a group of large deviation < threshold) 3. Resampling technique -> cluster has the same length -> take average to get the template waveform of each clusters 4. correlation of clusters, merge them if p > 0.9	Not stated clearly, but possibly: Atrial Fibrillation, Absence of P-wave, T-wave inverted Ischemic event or Myocardial Infarction	Extremely high Se(%) >= 99.97 P+ >= 99.99	Noise robust algorithm, Time comparison criteria: delta t <= 0.1 x RR-mean, Threshold for mean value of V-RR, Threshold for t-V and t-W, Threshold for mean of V-RR,

Tech	Features	Description	Disease	Accuracy	Note
		5. Detection of S* and Q* -> draw the strange line 6. Determine local extremes, maximums with highest distance to this line is the T and P wave 7. P wave absence will have cluster's length < 75% average 8. Calculate the trigonometric curve (abrupt change in the signal's slope), determine local maximums -> the offset and onset of P-wave and T-wave 9. With the T-offset, maximum, T-onset the time window of T-wave and P-wave templates are created 10. Correlation test with other waves -> highest correlation indicate the event of P-wave and T-wwave			P-wave absence: length < 75%, Trigonometric function G[n]
Time series analysis of **heart rate variability** (stochastic)	SDNN: standard deviation of the time of normal RR intervals (mils) SDAND: standard deviation of a mean of duration of RR intervals during each 5 minutes record RMSSD: square root of the mean of the squared of the differences between consecutive RR intervals pNN50: percentage of RR intervals that differ each other than 50ms	SDNN: the best statistical representation of cardiac mortality 3 years after MI Patients with SDNN < 70ms have 3-4 higher chance of death	Cardiac mortality after 3 years	Look into the article	

Tech	Features	Description	Disease	Accuracy	Note
Frequency analysis of Heart rate variability	Spectrum analysis of HRV: HRSA	HRVA evaluate the contribution of HRV on the autonomic nervous system Normal HRV consists of 3 dominant peaks: VLF: < 0.04Hz temperature regulation LF: 0.04 – 0.15Hz, sympathetic and parasympathetic activities HF: 0.15-0.4Hz, respiratory rhythm	Cardiac mortality after 3 years	Look into the article	Analysis of frequency usually associated with physiological perspective
Non-linear analysis of **Heart rate variability**	Power law exponent ***De-trended fluctuation analysis (DFA)*** ***Entropy***	Power law exponent: time series has similar fluctuation pattern with the frequency made up it. (from -1 to 1) DFA: similar to power law, but developed to distinguish between external and internal stimuli on the time series Entropy: measure the degree of randomness within a time series, greater value comes with greater disorder, evaluate heart rate dynamics	Cardiac mortality after 3 years	Look into the article	HR becomes more orderly with increasing age
Decision tree algorithm using ECG and BSPM	Abnormal ECG features on the 12 leads ECG (figure 5): STE, STD, Q wave, T inverse, LBBB, RBBB, LVH Body surface potential mapping variables regard ST and QRS duration: QRS width, axis, QRS and STT isointegrals, ST0 and ST60 isopotentials	12-ECG: STE based on the Minnesota code which requires 0.1 mV ST segment elevation in two or more of leads I, II, III, aVL, aVF, V5, V6 or 0.2 mV ST elevation in two or more of leads V1–V4 Body surface map diagnostic algorithm: Conduction delay was defined as epicardial	**Acute Myocardial Infarction** presented with confounders: LBBB, RBBB…	Physician interpretation of the results from the algorithm developed on BSPM criteria improves the detection of AMI (sensitivity 86%, specificity 98%)	Decision tree accomplished basing on some criteria on the acquired features

Tech	Features	Description	Disease	Accuracy	Note
		QRS duration 120msec, LBBB with AMI (see article), RBBB with AMI (see article), LVH and LVH with AMI (see integral)			
Transform of mono-polar ECG into **multichannel spectrum** domain	**f0**: frequency of the spectral peaks **w0**: its frequency bandwidth below 50% of the peak value **e0**: maximum Eigen value of the difference of the signal autocorrelation matrix **r0**: maximum difference in consecutive lags in the Autocorrelation sequence **Cj**: sum of squares of the first J reflection coefficients	Steps to obtain value f0, w0, e0, r0, Cj is described in the article. 1. Preprocessing: ECG sequence, subtract mean value, normalized by total energy, time a rectangular window -> final X(n) sequence 2. Generate spectrum: add zero padding, calculate FFT (S[k]), generate spectrum (S^2[k]), find max spectral component (S-max), determine max frequency (f0), find the bandwidth frequency below 50% of f0 (w0) -> enough for detection of ischemia 3. Autocorrelation sequence: generate this AC sequence, create AC matrix, compute Eigen value, Eigen-max, Eigen-differences sequence, AC difference sequence (r) and max of AC difference sequence (r0) 4. Run additional algorithm: Levinson-Durbin algorithm for AC sequence, compute CL, VL, EL parameters 5. Run statistical analysis on each of the parameters obtained, namely univariate analysis and multivariate analysis (combine e0, r0, w0) and validate technique accuracy using area under the ROC curve.	**Myocardial Ischemia**	Area under the ROC curve is given for each of the features: f0, w0, e0, r0, Cj and yield high sensitivity (>80%)	In the article, f0 and w0 are used to distinguish between ischemia and normal sinus. For ischemia f0 is << and shifted to the left. The probability of missing ischemia detection is 0.002 and probability of detecting normal condition is > 0.94
Wavelet Entropy Analysis	Wavelet Entropy:	High resolution ECG is obtained using **orthogonal leads** XYZ	**Ventricular tachycardia after**	Result: patients with LVP has:	HRECG is defined obtained with

Tech	Features	Description	Disease	Accuracy	Note
of High resolution ECG	from the peak of Q wave to end of QRS complex is calculated Calculated with **CWT** and **DWT**	Signal is then transformed using Continuous Wavelet Transform and Discrete Wavelet Transform, then applied with the entropy of the signal. **Wavelet entropy** is a function of time, represent the energy distribution within time-range -> can be used to analyze the disorder of the signal within specific time range In this study, the duration between R peak to QRS end point is studied to detect Ventricular Late Potential accompanied with Ventricular tachycardia after MI.	**Myocardial Infarction** reflected by the Late Ventricular Potential during the Q peak and QRS endpoint	**Higher disorder** (increasing, fluctuating entropy) **Lower Energy** (total area under the entropy curve) comparing to normal patients	XYZ leads, 1000Hz sampling rate with 12-bit data resolution
ECG-based Heart beat Classification	Various types of different technique for each steps is described. However, only the best will be named here for each: 1. Signal preprocessing: state-of-the-art classification paper [10] **does not even use preprocessing**, however, one worth mentioning is the FIR. 2. **Heart beat segmentation**: namely ***QRS detection***, using Pan & Tompkins algorithms 3. Feature extraction: most common is RR interval (fig 8) 4. **Classification**: **Reservoir Computing with Logistic Regression** (state-of-the-arc)	1. Signal preprocessing: FIR, wavelet transform, Bayesian filters for noise reduction, Extended Kaman filter, 2 median filter remove baseline wander, 2. Heart rate segmentation: Pan & Tompkins algorithm for QRS segmentation, neural networks [53], genetic algorithms [50], wavelet transform [60, 61, 4], filter banks [46], *Quad Level Vector* 3. Feature extraction: RR intervals has the famous features extracted (higher accuracy when normalized), nest is QRS interval, features extracted from wavelet transform (DWT and CWT) and VCG, then features from time-domain and frequency domain. Techniques used the reduce the number of features include: PCA, ICA (reduce the total of sample represent the heart beat), interpolation, Kernel PCA, clustering	**Arrhythmia Classification**	**Reservoir Computing (RC) has the highest, state-of-the-arc sensitivity**, suitable for real-time application and appropriate for computational cost: Sensitivity > 98% See figure 9	PCA perform better at noise removal, while ICA preforms best for extracting features They stressed that the most important features appears are RR intervals, the amplitude and length of the T wave, and 2nd-order statistics

Tech	Features	Description	Disease	Accuracy	Note
		technique, Generalized Discriminant Analysis (GDA), 4. Features selection: most important are RR intervals, T duration and amplitude and some 2-nd order statistic 5. **Learning algorithms**: Best 4 are Support Vector Machine (SVM), ANN, Linear Discriminant and Reservoir **Computing with Logistic Regression** (state-of-the-arc)			
Morphological interpretation of ST segment	**Morphological variables** about ST segment: ST slope, depth of T **Clinical variables**: heart rate, blood pressure **Others**: area-under-the-curve of the ST segment	- morphological characteristics of ST deviation: >= 1.0 mm from baseline, last for > 1 min - other variables: depth of depression, duration of the episode, area-under-the-curve of the ST-segment depression - clinical variables: **Heart rate**, RR-interval, VPB count, SVPB count - Limitation: + lack of data: 48h monitor, cost 40MB per records -> in sufficient data storage + reliance of detection algorithm + false positive due to changes in posture, rise in blood pressure,... can also cause STD - Good knowledge: + increase of heart rate, increase of **blood pressure** before STD is the current characteristics of silent ischemia + STD happen in **episodes**	**Silent Ischemia**	Accuracy = 64% with Specificity 67%	ST depression alone cannot diagnose silent Ischemia, usually coming along with increase in heart rate and blood pressure. STD happen in episodes Low sensitivity due to false STD STD accompany with long-term ECG is better than STD in stress test

Tech	Features	Description	Disease	Accuracy	Note
Smoothed De-trended Fluctuation Analysis (SDFA)	Calculate the **Hurst exponent (H)**	This article uses 2 different types of analysis technique + De-trended Fluctuation Analysis: + Wavelet Shrinkage: reduces the magnitude of terms in the high-pass portions. Finally, the wavelet transform is inverted to get the de-noised version of the data + Then calculate H: H<0.6 -> normal, H>= 0.6 -> arrhythmia	Arrhythmia Detection	Not stated	This is a typical example of applying stochastic method. How to calculate H, find in the article
Isoelectric Energy for Ischemic beat detection	Calculate the energy within a specific ST region Energy high -> closer to isoelectric line -> normal Energy low -> far from the isoelectric line -> ischemic	There are 5 processes: 1. Preprocessing: using wavelet transform to filter the signal, filter out also baseline wander, muscle electricity … 2. Delineation: detect R peak, then apply a threshold for detecting ST segment (RR/8, start at J point) 3. Calculation of **isoelectric energy** (equation in article) 4. Compare with threshold + validate that episodes of Ischemic beat lasts > 30s 5. Make conclusion	Ischemic beat	Sensitivity > 98%	Simple, applicable for real-time Can be potential variable for silent heart attack detection

METHODOLOGY

2.1 DATABASE ACQUISISTION

First, EKG physiological database dedicated for cardiovascular diseases detection is researched and documented. The target database needs to contain high quality, good resolution EKG signal with considerably long measuring time that is higher than couple of minutes. Therefore, the following databases used in this research all contain a decent signal acquisition system that is less subjected to noise, all come with 12 to 14 bits of data resolution, digitalized at 250 samples per second and the recording time ranges from 2 hours to 20 hours of continous monitoring. In addition, it is useful that clinical diagnosis is also documented to provide method for validation.

2.1.1 LONG-ST DATABASE

The Long-Term ST Database is the most novel EKG database dedicated for development and quantification of ischemia and other types of cardiovascular diseases. It contains 86 lengthy ECG recordings of 80 human subjects, chosen to exhibit a variety of events of ST segment changes, including ischemic ST episodes, axis-related non-ischemic ST episodes, episodes of slow ST level drift, and episodes containing mixtures of these phenomena. The greatest advantage using this database is that it also provides disease description for each patients, however, many of its records are subjected to high noise level that reder them unsuable for the scope of this research.

The following records are carefully chosen from the database, with the aim to only quantify clear signal while still provide various cases of myocardial injury.

```
data_path = 'C:\Nguyen Pham\MY
THESIS\database\longst\';
recordings = [20011 20021 20031 20041 20051 20061
20071 20081 20091 20101 20111 20121 20131 20141
20151 20171 20181 20191 20201 20211 20221 20231
20241 20251 20261 20271 20272 20274 20461 20161
20361];
leads = ones(1,length(recordings));
```

2.1.2 EUROPEAN DATABASE

Beside Long - ST database, the European database is also very famous in the field of medical application for cardiovascular damage. It is intended to be used for algorithms evaluation of ST and T-wave changes. This database consists of 90 annotated excerpts of ambulatory ECG recordings from 79 subjects. Not only that, Myocardial ischemia was diagnosed or suspected for each subject, additional selection criteria were established in order to obtain a representative selection of ECG abnormalities in the database, including baseline ST segment displacement resulting from conditions such as hypertension, ventricular dyskinesia, and effects of medication. Each of these records lasts for 20 hours and comes with very good quality and resolution.

The following records will be used in this research:

```
data_path = 'C:\Nguyen Pham\MY THESIS\database\euro\';
recordings = [103 104 105 106 112 113 118 121 129 133 136 139 154 161
162 163 170 105 108 112 115 123 129 133 147 154 104 112 118 122 154
161 612 801 808];
leads =      [001 002 001 002 002 002 001 001 002 002 002 002 002 002
002 002 001 001 001 002 002 002 002 002 002 002 001 002 002 002 001
002 002 002 002];
```

2.1.3 ST CHANGES DATABASE

Finally, ST changes database is a compact and suplementation to the European database. This database includes 28 ECG recordings of varying lengths, most of which were recorded during exercise stress tests and most of which exhibit transient ST depression. Due to this practice, most of the records have very high noise level, therefore, a rejection criteria need to be developed to reject outputs generated from noisy and unsuable input. This section will be covered later on in this section.

Because each record only lasts for 2 hours, all of them will be included in this research.

```
data_path = 'C:\Nguyen Pham\MY THESIS\database\stchange\';
recordings = 300:327;
leads = ones(1,28);
```

In conclusion, a total of 94 records will be used to train and validate the algorithm developed in this research. Each record has 12 or 14 bit data resolution, digitalized at 250 samples per second. The time range used for each records will be 1 hour long and a moving window of 10 seconds will be applied to calibrate the EKG features. The following code describes the whole process of reading EKG signal, signal preprocessing and baseline wander removal that will be covered in the next parts of this research. The research methodologies is also described in figure 15.

```
% filename: EURO_TASK.m
for record = 1:length(recordings)
    try
        filename = ['e0' num2str(recordings(record))];
        disp(filename);
        full_path = [data_path filename '.hea'];
        ECGw = ECGwrapper( 'recording_name', full_path);
        % READ SIGANL AND ANNOTATION-------------------------
        ann = ECGw.ECG_annotations;
        hea = ECGw.ECG_header;
        sig = ECGw.read_signal(1,hea.nsamp);
        sig1_raw = sig(:,leads(record));
        sig1_raw = sig1_raw(1:end);
        % NORMALIZATION CODES--------------------------------
        sig1_raw = sig1_raw - mean(sig1_raw);
        L = length(sig1_raw);
        Ex = 1/L * sum(abs(sig1_raw).^2);
        sig1_raw = sig1_raw / Ex;
        % BASELINE REMOVE USING Wavelet_decompose------------
        [approx, detail] = wavelet_decompose(sig1_raw, 8, 'db4');
        sig1 = sig1_raw - approx(:,8);
        sig_backup = sig1;
        % NORMALIZA THE SIGNAL FROM 0 TO 1
        sig1 = sig1 + abs(min(sig1));
        sig1 = sig1 / max(sig1);
        % GENERAL PARAMETERS---------------------------------
        fs = hea.freq;
        ts = 1/fs;
        try
            REPORT;
        catch
            disp('An error occured while calibrating this record');
            failed_records{end + 1} = filename;
        end;
    catch
        disp(['record ' filename ' not found. Proceed to next one']);
    end;
end;
```

2.2 SIGNAL PREPROCESSING

2.2.1 SIGNAL NORMALIZATION

The purpose of normalization is to scale down the signal into a specific range so that records from different databases become comparable to each other. There are various techniques for signal normalization. In this research, the signal is normaized against its net energy.

1. Firstly, the signal is substracted by its mean

$$signal = signal - mean(signal)$$

2. Then net energy value is calculated as follow:

$$net = \frac{1}{length(signal)} * \sum abs(signal)^2$$

3. Devide the orginal signal with the net energy to get the scaled version

$$signal = signal - net$$

4. Then change the scale into from 0 to 1 with the equation

$$signal = \frac{\left(signal + abs(\min(signal))\right)}{\max(signal)}$$

Matlab code presentation:

```
% NORMALIZATION CODES--------------------
    sig1_raw = sig1_raw - mean(sig1_raw);
    L = length(sig1_raw);
    Ex = 1/L * sum(abs(sig1_raw).^2);
    sig1_raw = sig1_raw / Ex;
% NORMALIZA THE SIGNAL FROM 0 TO 1
    sig1 = sig1 + abs(min(sig1));
    sig1 = sig1 / max(sig1);
```

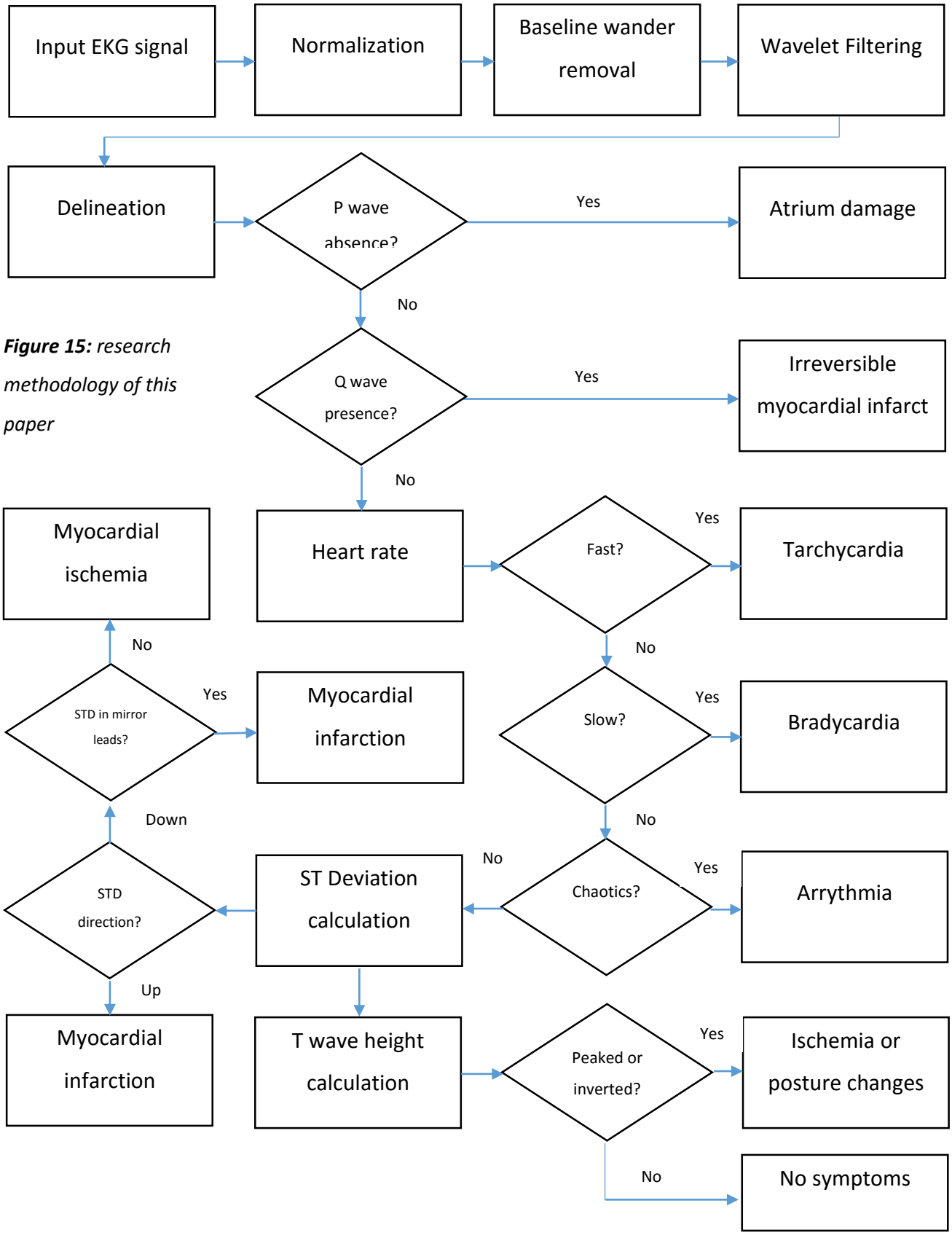


Figure 15: *research methodology of this paper*

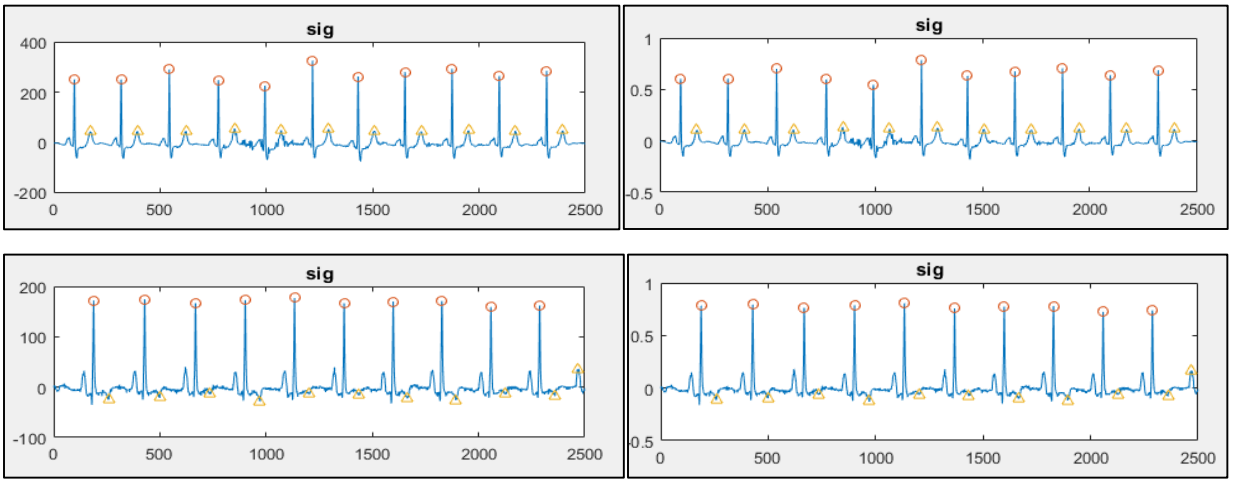


Figure 16: *Result obtained after normalizing the signal. Two segment of data with different value ranges are scaled down into the range of from 0 to 1 to make comparison between them becomes feasible.*

2.2.2 BASELINE WANDER REMOVAL

During recent years, many researchers have constantly report the use of wavelet decomposition for removing low frequency components within biological signals. According to Khawala, the baseline wander of EKG signal can be accurately located at the 9^{th} level approximation coefficient of Daubechies11 mother wavelet. In another study, the authors also suggest using Daubechies4 mother wavelet to perform up to 4^{th} level of decomposition and use the approximation coefficient at level 4^{th} to address the baseline wander. Baseline wander removal using wavelet decomposition is one of the most novel and state of the art in the field of EKG signal processing because it is not only effective, but also address the disadvantages corresponding to the conventional FIR and IIR filter design. It is very well known that after applying FIR or IIR filter into EKG signal, the morphology of ST segment will be altered, which will have critically negative effect on the analysis of the signal because ST segment is the most important feature for addressing many types of cardiovascular diseases.

The table below is extracted from Khawala's paper, in which he performed baseline wander removal using many different types of mother wavelet. From his research, it is documented that the best technique for removing baseline wander is using Daubechies11 mother wavelet to perform signal

decomposition. The approximation coefficient at the 11th level will be chosen to be the baseline wander and it will be substracted from the orginal signal.

Order No.	Mother Wavelet	n	mean $CE\%$	mean $CB\%$
1	db11	9	99.9924	99.9150%
2	sym12	9	99.9913	99.9011%
3	sym10	9	99.9909	99.8962%
4	db10	9	99.9906	99.8925%
5	coif5	9	99.9904	99.8894%

***Figure 17**: Result obtained from Khawala research. The best mother wavelet for removing baseline wander for EKG signal is Daubechies11 wavelet at decoposition level 11th*

Matlab code presentation:

```
%-BASELINE REMOVAL-------------------------------------
 [approx, detail] = wavelet_decompose(seg, 11, 'db11');
 seg = seg - approx(:,11);
 baseline = approx(:,11);
```

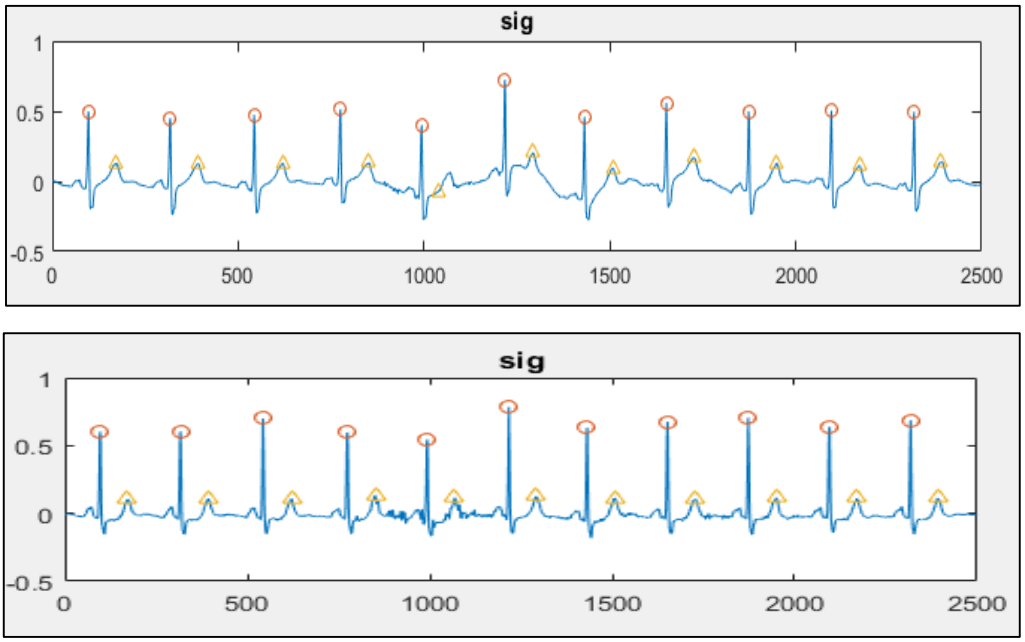


***Figure 18**: Daubechies11 mother wavelet successfully remove the effect of baseline wander on the EKG signal*

2.2.3 ESTIMATION AND REMOVAL OF NOISE AND MUSCLE ACTIVITIES

Noise from the EKG signal comes from mnay different sources. Among these, muscle activities are the most prominent source of noise that greatly affect the quality of detection algorithm. EKG signal processing has been tremendously studied to determine the best frequency range that have critiical effect in removing muscle activities while retaining the morphologies of the chracteristic waves. It has been shown that appropriate frequency range that could be used for removing muscle activities is from 0.5Hz to 50Hz. In other researches, Discrete Wavelet Transform (DWT) is also very well known for its ability to remove noisy components from EKG signal that could represent for muscle acitivites. Further documentation about this novel technique can be found in Khawala's research. After the process of noise removal, a rejection criteria will also be developed to remove the remaining extremely noisy segment because the technique can not absolutely remove all sources of noise. The following steps will be applied to obtain noise free signal that will be used for algorithm development.

1. First, a bandpass filter using FIR design technique, with a cutoff frequency ranges from 0.5 to 50Hz, filter order of 24 and a rectangle moving window will be designed.
2. The filter is convoluted with the original signal to obtain the first level noise free signal.
3. The signal will then be decomposed using Discrete Wavelet Transform (DWT), with a Sym2 mother wavetlet following 1 level of decomposition. The first level detail coefficient (CD) will be chosen to represent the noise from muscle activities.
4. Establishing a threshold for CD with the following criteria

$$CD = \begin{cases} CD\ with\ abs(CD) > Threshold \\ 0\ with\ abs(CD) \leq Threshold \end{cases}$$

5. The theshold can be chosen as fixed, with some information gathered prior to the signal itself. Under the assumption of white noise, the threshold with ω variance, as described in Khawala research, can be chosen as follow:

$$Threshold = \omega \sqrt{2.\ln(length(signal))}$$

6. Since ω is often unknown in practice, it is estimated as the medican of the absolute deviation which avoids the influence of the outliner values.

$$\omega = 1.483\ median(CD)$$

Matlab code presentation:

```
%-SIGNAL PREPROCESSING-------------------------------------------
%-BASELINE REMOVAL-----------------------------------------------
[approx, detail] = wavelet_decompose(seg, 11, 'db11');
seg = seg - approx(:,11);
baseline = approx(:,11);
%-NOISE ESTIMATION REMOVAL---------------------------------------
[approx, detail] = wavelet_decompose(seg, 1, 'sym2');
noise_level = detail(:,1);
noise_variance = 1.483 * median(noise_level);
noise_threshold = noise_variance * sqrt(2 * log(length(noise_level)));
  for hkn = 1:length(noise_level)
      if abs(noise_level(hkn)) < noise_threshold
         noise_level(hkn) = 0;
      end;
  end;
%-DENOISING------------------------------------------------------
seg = seg - noise_level;
```

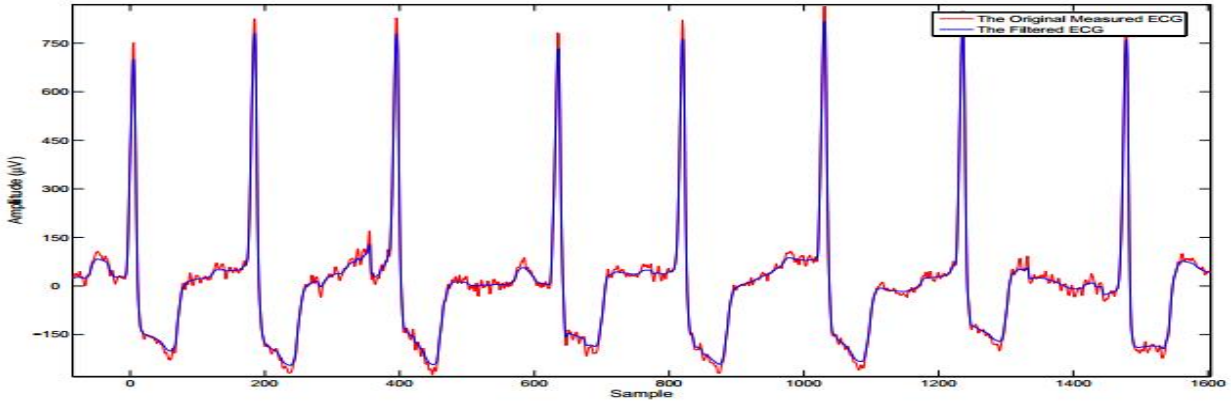


Figure 19: *Comparison between the noisy signal and filtered signal obtained from the previous denoising technique*

2.3 EKG DELINEATION

The purpose of EKG delineation is capturing the most prominent peaks and event of the characteristic waves, including P wave, QRS complex, R peak, ST segment onset and offset and T wave. The interest of this research is performing R peak detection, T wave detection and ST segment detection by specifing ST onset and offset.

2.3.1 R PEAK DETECTION

R peak detection of EKG signal has been the subject of interest for many decades..In other to develop analysis algorithm using EKG signal, it is extremely important that the dection of R peak need to be extremely precise. There are plenty of traditional and novel method for detection of R peak detection, including Pan Tompkins algorithm or Kaman filter, Neural Network and Wavelet based decompositon technique. In this research, a new detection algorithm will be described and the accuracy of the method will be compared with other methods in the final thesis.

1. First, ECG segment will be stored in another variable dedicated for transformation and filtering
2. The segment will be filtered using a bandpass filter, designed using FIR filter design, with a frequency range from 10Hz to 25Hz to capture the QRS complex while removing other waveforms
3. The whole segment will be shifted above 0 before squaring, so that the prominent peak obtained is only represented for the R peak because there is no minima that is located below the zero baseline that could becomes prominent after signal quaring. This step is the main different of this technique comparing to the Pan Tompkins algorithm.
4. Instead of squaring, the signal is powered by 12 to maximize the effect of the R peaks on the signal.
5. Next, local maxima will be calculated
6. A moving window of 200ms will be applied to scan the whole signal. A value is chosen to be an R peak if it qualifies all of the following criteria:
    - It is a local maxima

- It is larger than the threshold
- It locates at least 300ms behind the first highest peaks

With the threshold is chosen as the mean value of the segment in addition to its standard deviation.

$$Threshold = mean(segment) + std(segment)$$

7. Then, the location of these values are looked up within the original signal. The amplitude and location will then be stored in the coresponsing variables for further analysis. Figure 20 describles the quality of this R peak and T peak detection algorithm.

Matlab code presentation:

```
%-READ SIGNAL-----------------------------------------------------------
data_length = span * fs;
startpoint = (inputloop - 1) * data_length + 1;
endpoint = inputloop * data_length;
seg = sig1(startpoint:endpoint);
% GENERAL PARAMETERS-----------------------------------------------
QRS_amps = [];
QRS_locs = [];
QRS_amps2 = [];
QRS_locs2 = [];
T_amps = [];
T_locs = [];
%-QRS DETECTION---------------------------------------------------------
%-Filter 10 - 25Hz to remove other waves---------------------------
filt = fir1(24, [10/(fs/2) 25/(fs/2)],'bandpass');
seg2 = conv(filt,seg);
seg2 = seg2(12:end,1);
%-BRING THE SIGNAL ABOVE 0----------------------------------------
seg2 = seg2 + abs(min(seg2));
seg2 = seg2.^12;
thres_mean = (mean(seg2) + QRS_std_thres * std(seg2)) * ones(1,length(seg2));
 [pks, locs] = findpeaks(seg2,'MinPeakDistance',100);
for i = 1:length(locs)
    if pks(i) > thres_mean(1)
       ind = locs(i);
       QRS_amps(end + 1) = seg(ind);
       QRS_locs(end + 1) = ind;
    end;
end;
```

2.3.2 T PEAK DETECTION

This section represents a simple yet effective technique to locate the T peaks after performing R peak detection.

1. The first step is capturing the isoelectric baseline of the signal. The value of isoelectric baseline is chosen to be the value located at the middle of the RR interval

$$isoelectric = signal(0.5 * length(RR))$$

2. Next step is performing baseline transformation to each RR interval using this criteria

$$RR(i) = \begin{cases} RR(i)\ if\ 0.2 * length(RR) \leq i\ \leq 0.6 * length(RR) \\ 0\ otherwise \end{cases}$$

The purpose of this step is to consider only the part of the RR segment where the T peak can take place. The interval is chosen to be between 0.2 to 0.6 of the RR interval

3. The segment is then subtracted with the isoelectric baseline and then powered by 12 to capture the maximum effect of peaks within this segment. The process of substracting isolectric baseline is crucial because T wave can both be tranverse or inverse.

$$RR(i) = (RR(i) - isoelectric)^{12}$$

4. T wave is then detected as the highest value of this segment

$$Tpeak_{location} = \max(RR(i))$$

Matlab code presentation:

```
%-T WAVE DETECTION-------------------------------------------------------
sig = seg;
%-ZEROING EACH BEAT INTERVAL---------------------------------------------
for hk = 1:length(QRS_locs) - 1
    data = sig(QRS_locs(hk):QRS_locs(hk + 1));
    leng = floor((QRS_locs(hk + 1) - QRS_locs(hk)) / 10);
    isoelec = seg(QRS_locs(hk) + 5 * leng);
    data = data - isoelec;
    data = abs(data);
    data = data.^12;
    data = data(2 * leng:6 * leng);
    [val, ind] = max(data);
    T_locs(end + 1) = QRS_locs(hk) + 2 * leng + ind;
    T_amps(end + 1) = seg(QRS_locs(hk) + 2 * leng + ind);
end;
```

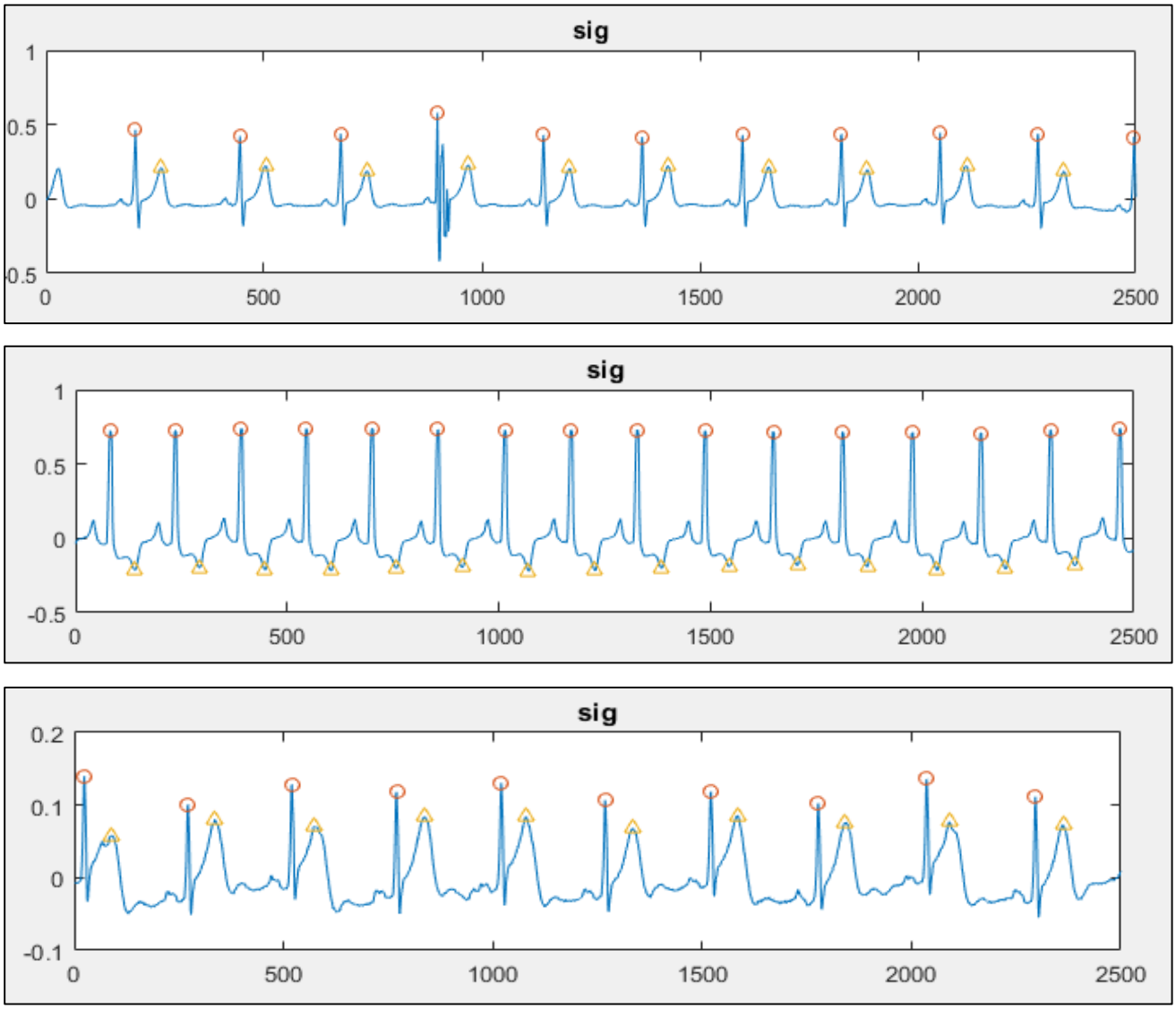


Figure 20: *Result obtained after performing EKG delineation. The location of R peaks and T peaks are correctly determined.*

2.3.3 REJECTION CRITERIA

In some cases the signal is extremely noisy that can not be used for algorithm development process. Therefore after the EKG delineation takes place, it is important to reject the segment with unwanted results. The segment will be rejected if one of the following criteria is met:

1. Any R peak and its corresponding T peak is located too far away (larger than 1000ms) or too close (less than 25ms)
2. Any consecutive R peaks that is located too far away (larger than 2000ms) or too close (less than 100ms)
3. Any consecutive T peaks that is located too far away (larger than 2000ms) or too close (less than 100ms)

Matlab code presentation:

```
% REJECTION CRITERIA-------------------------------------------------
        for rjloop = 2:length(QRS_locs)
            condition = QRS_locs(rjloop) - QRS_locs(rjloop - 1);
            if condition > 450
                rejected = rejected + 1;
                continue;
            end;
        end;
        for rjloop = 2:length(T_locs)
            condition = T_locs(rjloop) - T_locs(rjloop - 1);
            if condition > 450
                rejected = rejected + 1;
                continue;
            end;
        end;
        for rjloop = 1:length(T_locs)
            condition = T_locs(rjloop) - QRS_locs(rjloop);
            if condition > 250 || condition < 25
                rejected = rejected + 1;
                continue;
            end;
        end;
```

2.3.4 ST SEGMENT DETECTION

After detecting R peaks and T peaks, ST segment is described as the interval from 0.4 from 0.65 of the RT interval. ST segment is crucial for calibrating ST devication and ST slope, which are the most important feaures that have been tremendously cultered in the field of myocardial injury analysis using EKG signal.

Matlab code presentation:

```
for km = beat_start:beat_end
        if ~isnan(QRS_locs(km)) && ~isnan(T_locs(km))
            leng = floor((T_locs(km) - QRS_locs(km))/4);
            pheight = (seg(QRS_locs(km) + floor(2.6 * leng)) -
                  seg(QRS_locs(km) + floor(1.6 * leng))) / seg(QRS_locs(km)) *
                  100;
            width = floor(2.6 * leng) - floor(1.6 * leng);
            STslope(end + 1) = pheight / width * 10;
            Tinv(end + 1) = seg(T_locs(km));
            ToR(end + 1) = abs(seg(T_locs(km))) / abs(seg(QRS_locs(km))) *
                  100;
            ST_on_locs(end + 1) = QRS_locs(km) + floor(1.6 * leng);
            ST_on_amps(end + 1) = seg(QRS_locs(km) + floor(1.6 * leng));
            ST_off_locs(end + 1) = QRS_locs(km) + floor(2.6 * leng);
            ST_off_amps(end + 1) = seg(QRS_locs(km) + floor(2.6 * leng));
        end;
    end;
```

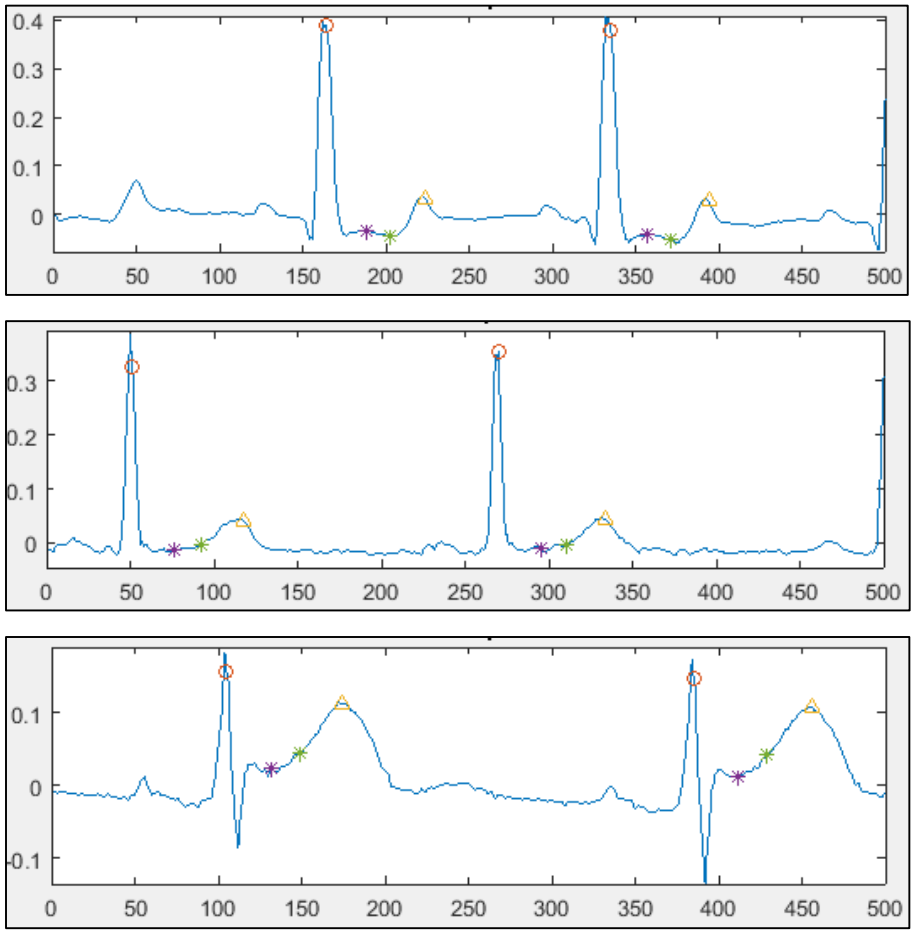


Figure 21: *ST segment detection quality*

2.4 FEATURES EXTRACTION

Reminding that features extraction will be performed for each data segment lasting for 10 seconds. Firstly, the following code breaks down the entire signal into different segments, then features extraction is performed for each beat located within each segment. The feature values representing for each segment are taken as the mean of the consecutive value calibrated during each beat. The following code describes the process of calibrating number of loops for each record as well as looping the entire signal length.

Matlab code presentation:

```
fraction = 1/2;
% CALCULATE SOLOOP--------------------------------------
total_length = length(sig1);
window_length = fs * span;
number_of_loop = floor(total_length * fraction /
window_length);
for soloop = 1:number_of_loop
   do something …
end;
```

2.4.1 MORPHOLOGICAL FEATURES

In other to detect abnormalities within EKG signal that represents for myocardial infarction and ischemia, detection of ST deviation and ST slope need to be accomplised. Revise literature review that if there is a transient ST deviation and ST slope upwarding, the patient is prescripted with ST elevation myocardial infarction. If ST deviation and ST slope are downwarding, the patient is suspected with ST depression myocardial infarction if the mirror leads have transient symptoms of ST deviation and ST slope upwarding, or prescribed with Ischemia if such abnormalities do not occur. In case of no transient ST segment elevation or depression, the direction of T wave is considered. T wave is essentially upward in most of EKG leads and often exhibit downwarding characteristics or even becomes significantly peaked when ischemic episode occurs. However, recent researches have shown that T wave downwarding can be a result from posture changes that does not necessarily

indicate cardiovascular injury. Therefore, the clinical value of T wave direction and amplitude is not as sensitive as the ST slope and ST Deviation, but it is also necessary to be calculated.

Therefore, up to 4 morphological features will be calibrated within the scope of this research. These are ST Deviation, ST slope, T direction and T amplitude. The methodology is described below:

1. ST deviation is calculated as the area under the curve of ST segment as derived from section 2.3.4 and the isoelectric baseline. Then the value is normalized with the length of the segment to calculate the imerical height that represent the deviation from the isoelectric line.
2. ST slope is calibrated as the tan value between the line connect the onset and offset of ST segment againts the isoelectric line.
3. T wave derection is calculated as the different between the amplitude of T peak with the isoelectric line.
4. T amplitude score is calculated as the division of T amplitude againts the R peak amplitude, then scaled to persentage value.

Matlab code presentation:

```
%-CALDULATE Stslope, Tinv and ToR-------------------------------------
------
    STslope = [];
    for km = beat_start:beat_end
        if ~isnan(QRS_locs(km)) && ~isnan(T_locs(km))
            leng = floor((T_locs(km) - QRS_locs(km))/4);
            pheight = (seg(QRS_locs(km) + floor(2.6 * leng)) -
seg(QRS_locs(km) + floor(1.6 * leng))) / seg(QRS_locs(km)) * 100;
            width = floor(2.6 * leng) - floor(1.6 * leng);
            STslope(end + 1) = pheight / width * 10;
            Tinv(end + 1) = seg(T_locs(km));
            ToR(end + 1) = abs(seg(T_locs(km))) / abs(seg(QRS_locs(km)))
* 100;
            ST_on_locs(end + 1) = QRS_locs(km) + floor(1.6 * leng);
            ST_on_amps(end + 1) = seg(QRS_locs(km) + floor(1.6 * leng));
            ST_off_locs(end + 1) = QRS_locs(km) + floor(2.6 * leng);
            ST_off_amps(end + 1) = seg(QRS_locs(km) + floor(2.6 *
leng));
        end;
    end;
```

```
%-CALDULATE STDeviation-------------------------------------------------
    for km = beat_start:beat_end
        if ~isnan(QRS_locs(km)) && ~isnan(T_locs(km))
            leng = floor((T_locs(km) - QRS_locs(km))/4);
            pdata = seg((QRS_locs(km) + floor(1.6 * leng)):(QRS_locs(km)
+ floor(2.6 * leng)));
            RRinterval = QRS_locs(km + 1) - QRS_locs(km);
            iso = ones(length(pdata),1) * seg(QRS_locs(km) + floor(0.5 *
RRinterval));
            STDeviation(end + 1) = (trapz(pdata) - trapz(iso)) /
length(pdata) * 100 * 10;
        end;
    end;
```

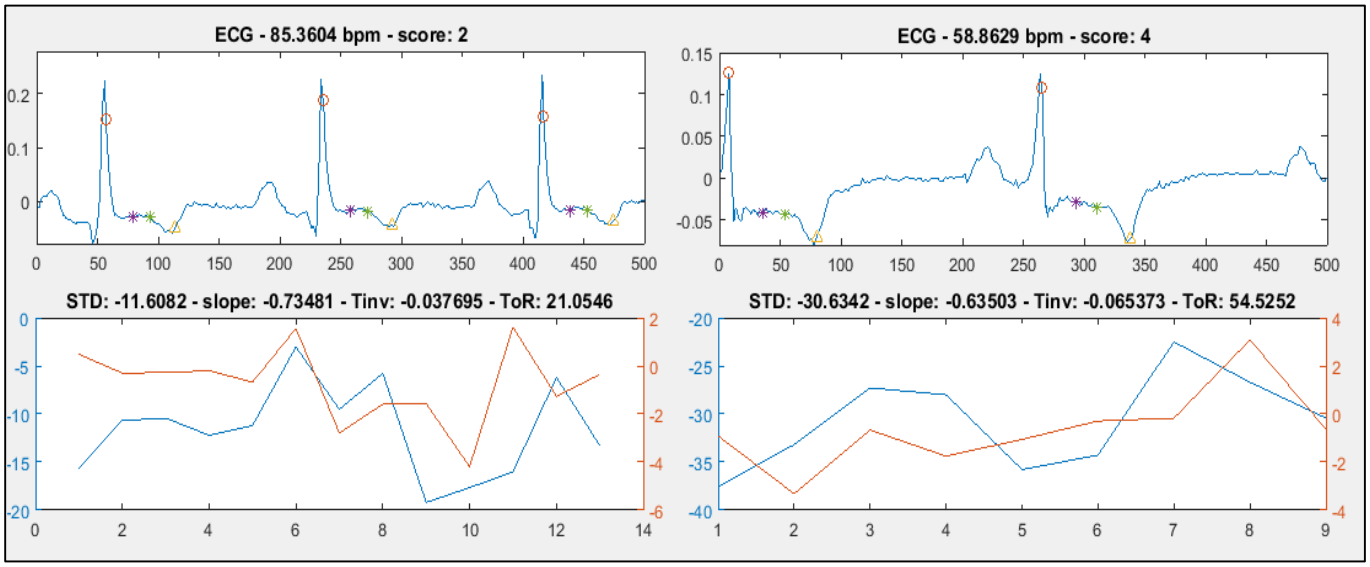


Figure 22: *Morphological features obtained after calculation. The consecutive values are presented as a graph and the mean value is displayed in the title of each graph.*

2.4.2 INTRA - BEAT DETRENDED FLUCTUATION

During many recent studies, it has been shown that Detrended Fluctuation Analysis had demonstate its potential as a feature for detection of cardiovascular diseases. Detrended Fluctuation Analysis analyzes the chaotic behavior of a system without considering its trend. First, the segment is devided into different box of data. Then a polynomial fucntion is fit into the data in each box, representing

the trend. After that, the signal within each box is substracted with its trend and the degree of fluctuation is calculated as:

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_n(k)]^2}$$

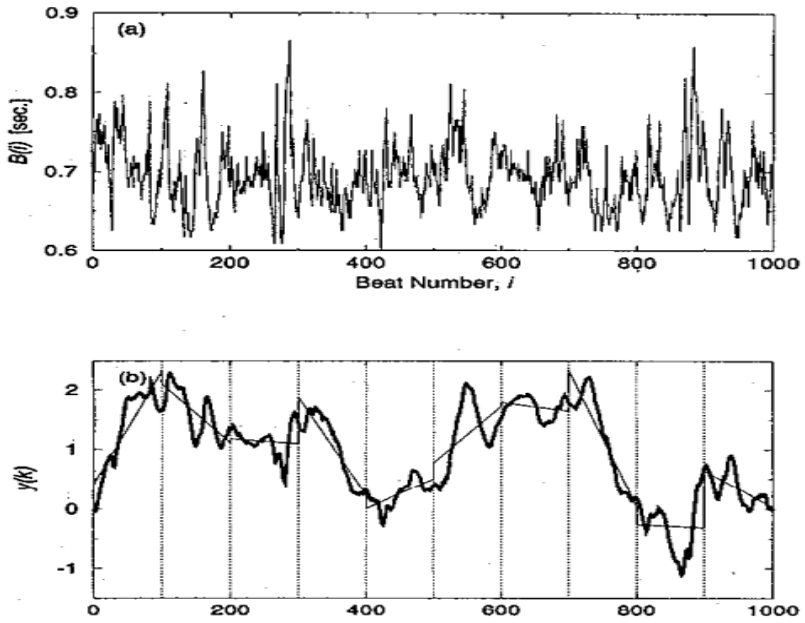


***Figure 22**: The data is divided into several boxes and the trend within each box is computed*

The computation is repeated for all of the boxes to provide the relationship between the fluctuation $F(n)$and the box size. A linear relationship on the double log graph indicates the present of scaling and the value of scaling exponent $\propto$ is calculated as the slope of line relating $\log(F(n))$ to $\log(n)$. According to many researchs, $\propto$ will be higher than 1 for patients who are suffering from coronary artery dieases, while staying mostly below 1 for healthy subjects.

In this research, each RR interval will be fetched into a matlab fucntion to calculate the value of $\propto$ for each beat. Then the DFA feature is defined as the mean value of all consecutive values. The purpose of this pratice is to quantify the sensitivity and specificity of DFA in detecting cardiovascular damage, finding the answer for the questions of what type of disease it represent for and how accurate it is in detecting of this disease.

Matlab code presentation:

```
%-CALCULATE DFA-------------------------------------------------
    for i = beat_start:beat_end
       data = seg(QRS_locs(i):QRS_locs(i+1));
       dfa = DetrendedFluctuation(data);
       DFA(end + 1) = dfa;
    end;
   DFA = DFA';
    RP_DFA_bin = [RP_DFA_bin; mean(DFA)];
```

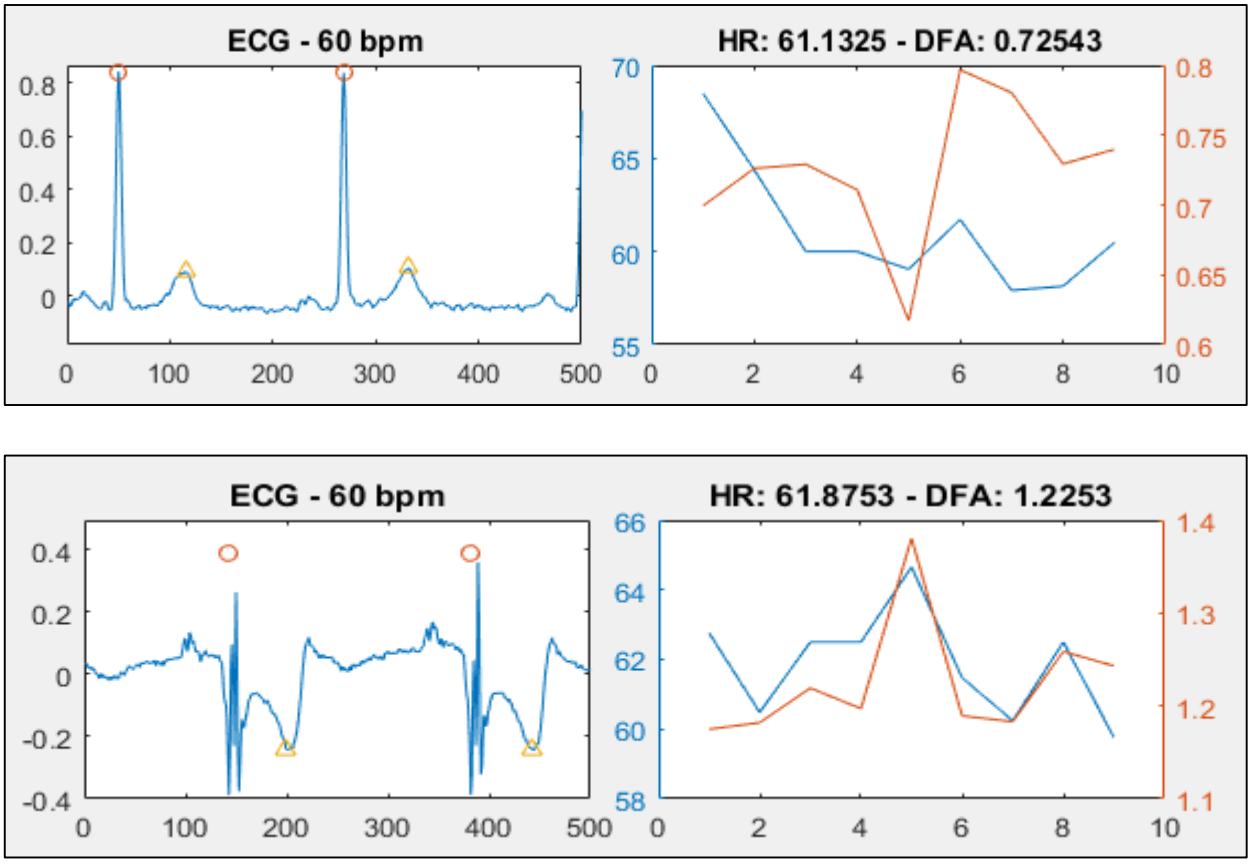


Figure 22: *DFA value obtained for each data window of length 10 seconds*

2.5 SCORING SYSTEM

Untill this part, up to 5 features have been extracted for the analysis of Myocardial infaction and ischemia. Using these 5 features, a scoring system will be developed to address the type and level of injury for each data window. If one of the following criteria is met, the overall score is increased by 1.

1. Transient ST deviation value that is higher than a maximum threshold or lower than a minimum threshold
2. Transient ST slope upwarding or downwarding value that is higher than a maximum threshold or lower than a minimum threshold
3. Presence of T wave inversion or T wave peaked
4. Presence of DFA value greater than 1

Matab code presentation:

```
if mean_STD > 20 || mean_STD < -20
        score = score + 1;
    end;
    if mean_STS > 8 || mean_STS < 0
        score = score + 1;
    end;
    if mean_Tinv < 0.02
        score = score + 1;
    end;
    %if mean_HR > 100 || mean_HR < 50
    %    score = score + 1;
    %end;
    if mean_DFA > 1
        score = score + 1;
    end;
```

The threshold for each criteria above is chosen intuitively during this state. In the future work, it will be compared to actual diagnosis from doctors and physician for better qualification. The process will be covered in the final thesis. The following code provides extra general clinical prescription for the data segment.

```
%-MAKING THE DIAGNOSIS-------------------------------------
---------
    if mean_STD > 100 || mean_STS > 11
        diagnosis = 'Transient ST Elevate';
    elseif mean_STD < -40 && mean_STS < -4
        diagnosis = 'Transient ST Depress';
    elseif mean_Tinv < -0
        diagnosis = 'T wave inverted';
    elseif mean_Tinv < 0.02
        diagnosis = 'T wave absence';
    elseif mean_DFA > 1 && mean_STD > 20
        diagnosis = 'Minor positive STD';
    elseif mean_DFA > 1 && mean_STD < -10
        diagnosis = 'Minor negative STD';
    elseif mean_STD > 20
        diagnosis = 'Minor positive STD without DFA';
    elseif mean_STD < -10
        diagnosis = 'Minor negative STD without DFA';
    elseif mean_DFA > 1
        diagnosis = 'STD spotted by DFA without MF';
    else
        diagnosis = 'Normal ECG';
    end;
```

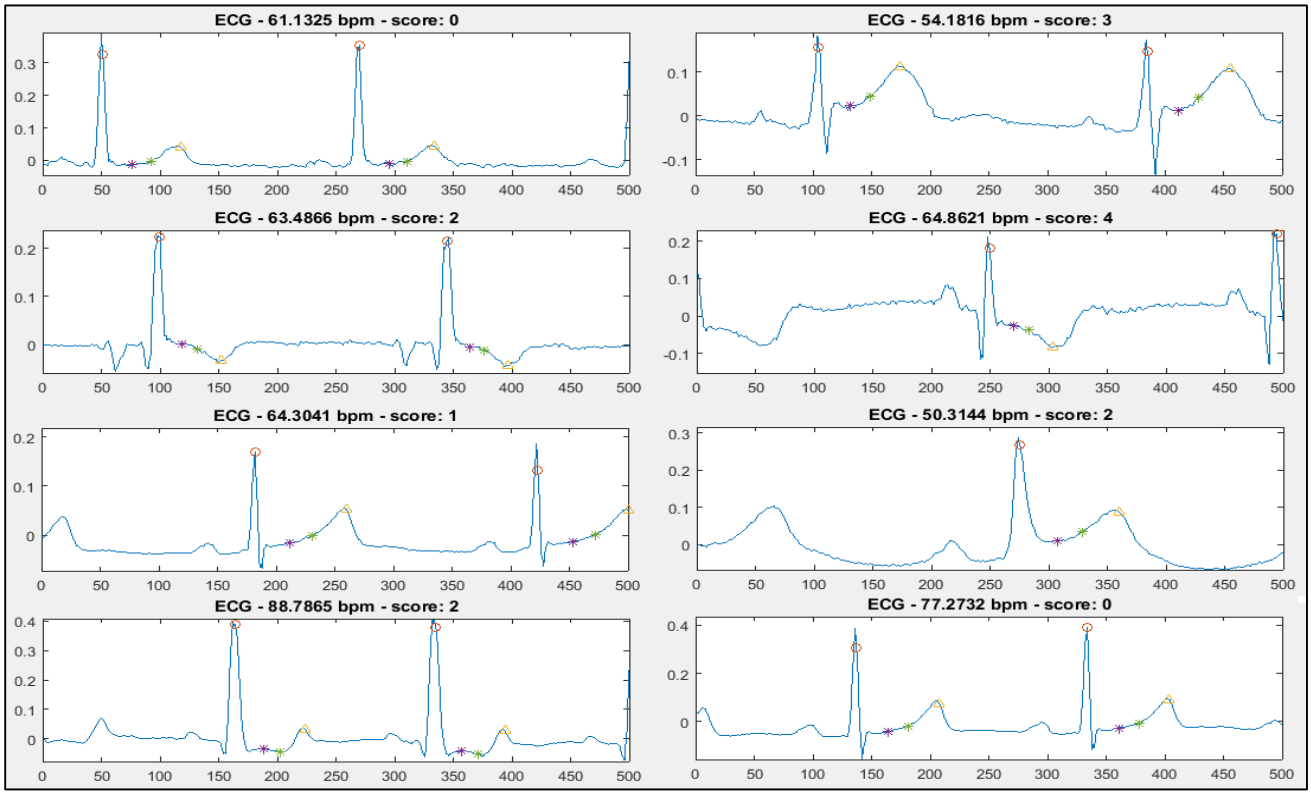


***Figure 23**: Scoring system creates the risk score for each of the data segment*

RESULT AND VALIDATION

In this section, the results obtained from performing features extraction and the potential quality of the risk score system in diagnosting vardiovascular damage will be covered.

3.1 FEATURES EXTRACTION SUMMARY

After performing feature extraction for each window length of 10 seconds, a package of total 5 features are extracted: ST deviation, ST slope, T wave direction T wave amplitude and DFA score. The process is repeated for all 94 records described in section 2.1, each record contains the data that last for approximately 60 minutes. In total, a table of 31,000+ rows are automatically computed to provide the database for validation and further analysis. The table below summarize the descriptive statistics of this database:

SCORE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	7894	24.9	24.9	24.9
	Caution	7710	24.3	24.3	49.3
	Risk	6796	21.5	21.5	70.7
	Danger 1	7152	22.6	22.6	93.3
	Danger 2	2119	6.7	6.7	100.0
	Total	31671	100.0	100.0	

Figure 24: *Descriptive statistic for the SCORE system, with score = 0 indicates Normal, score = 1 indicates Caution, score = 2 indicates Risk and score > 3 indicates Danger.*

As we can clearly see from the descriptive statistic table, the number of cases among the groups are approximately similar, each accompanies about 22% of the total cases. This is the result of careful and manual selection of different types of diseases observed within the first few minutes of the record. The final class (Danger 2), however, lacks the appropriate number of cases comparing to other classes. This class represent for the myocardial ischemia disease because it represents the highest possible score where ST deviation, ST slope downward and T wave inversion happen at the same time. The records representing for this disease were carefully chosen, however, during the period of 1 hour calculation, the EKG signal may have altered and the obtained is the increase in other classes while

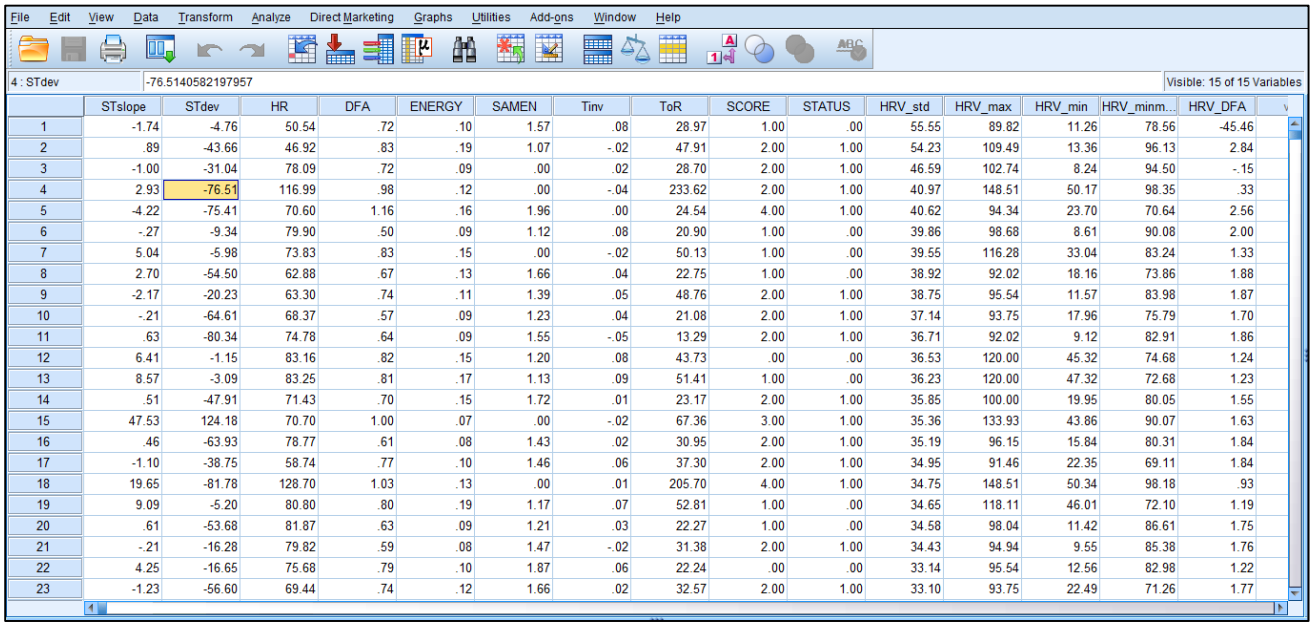
the number of cases for this classes is not met. Further effort will be made in order to equalize the number of cases within each classes for the further researches.

STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	15604	49.3	49.3	49.3
	ST Deviation	16067	50.7	50.7	100.0
	Total	31671	100.0	100.0	

Figure 25: *Descriptive statistic for the STATUS variable, in which ''Normal'' subject has the score of less than 2 and the ''ST Deviated'' subject has the otherwise.*

It is always a good practice to create a database with appropriately equal number of cases between each class to avoid the effect of overfitting and underfitting.



	STslope	STdev	HR	DFA	ENERGY	SAMEN	Tinv	ToR	SCORE	STATUS	HRV_std	HRV_max	HRV_min	HRV_minm...	HRV_DFA
1	-1.74	-4.76	50.54	.72	.10	1.57	.08	28.97	1.00	.00	55.55	89.82	11.26	78.56	-45.46
2	.89	-43.66	46.92	.83	.19	1.07	-.02	47.91	2.00	1.00	54.23	109.49	13.36	96.13	2.84
3	-1.00	-31.04	78.09	.72	.09	.00	.02	28.70	2.00	1.00	46.59	102.74	8.24	94.50	-.15
4	2.93	-76.51	116.99	.98	.12	.00	-.04	233.62	2.00	1.00	40.97	148.51	50.17	98.35	.33
5	-4.22	-75.41	70.60	1.16	.16	1.96	.00	24.54	4.00	1.00	40.62	94.34	23.70	70.64	2.56
6	-.27	-9.34	79.90	.50	.09	1.12	.08	20.90	1.00	.00	39.86	98.68	8.61	90.08	2.00
7	5.04	-5.98	73.83	.83	.15	.00	-.02	50.13	1.00	.00	39.55	116.28	33.04	83.24	1.33
8	2.70	-54.50	62.88	.67	.13	1.66	.04	22.75	1.00	.00	38.92	92.02	18.16	73.86	1.88
9	-2.17	-20.23	63.30	.74	.11	1.39	.05	48.76	2.00	1.00	38.75	95.54	11.57	83.98	1.87
10	-.21	-64.61	68.37	.57	.09	1.23	.04	21.08	2.00	1.00	37.14	93.75	17.96	75.79	1.70
11	.63	-80.34	74.78	.64	.09	1.55	-.05	13.29	2.00	1.00	36.71	92.02	9.12	82.91	1.86
12	6.41	-1.15	83.16	.82	.15	1.20	.08	43.73	.00	.00	36.53	120.00	45.32	74.68	1.24
13	8.57	-3.09	83.25	.81	.17	1.13	.09	51.41	1.00	.00	36.23	120.00	47.32	72.68	1.23
14	.51	-47.91	71.43	.70	.15	1.72	.01	23.17	2.00	1.00	35.85	100.00	19.95	80.05	1.55
15	47.53	124.18	70.70	1.00	.07	.00	-.02	67.36	3.00	1.00	35.36	133.93	43.86	90.07	1.63
16	.46	-63.93	78.77	.61	.08	1.43	.02	30.95	2.00	1.00	35.19	96.15	15.84	80.31	1.84
17	-1.10	-38.75	58.74	.77	.10	1.46	.06	37.30	2.00	1.00	34.95	91.46	22.35	69.11	1.84
18	19.65	-81.78	128.70	1.03	.13	.00	.01	205.70	4.00	1.00	34.75	148.51	50.34	98.18	.93
19	9.09	-5.20	80.80	.80	.19	1.17	.07	52.81	1.00	.00	34.65	118.11	46.01	72.10	1.19
20	.61	-53.68	81.87	.63	.09	1.21	.03	22.27	1.00	.00	34.58	98.04	11.42	86.61	1.75
21	-.21	-16.28	79.82	.59	.08	1.47	-.02	31.38	2.00	1.00	34.43	94.94	9.55	85.38	1.76
22	4.25	-16.65	75.68	.79	.10	1.87	.06	22.24	.00	.00	33.14	95.54	12.56	82.98	1.22
23	-1.23	-56.60	69.44	.74	.12	1.66	.02	32.57	2.00	1.00	33.10	93.75	22.49	71.26	1.77

Figure 26: *The analysis is carried on SPSS statistics. The table above contains in total 31,761 rows of data*

3.2 ACCURACY VADIATION

3.2.1 MORPHOLOGICAL FEATURES

Any abnormalities observed within a morphological feature is defined as the transend of this value over a defined threshold. At this state, the treshold is intuitively specified according to the author's knowledge about EKG clinical interpretation. In future work, this process will be qualified by doctors and clinican for better quatification of the thresholds for each features. However, it is observable that the risk score value calculated from morpholocal features is currently appropriate in both describing and classifying cardiovas cular diseases. The following table describes the theory of classification.

		Score	Description	Disease types
Valid	Normal	0	Normal EKG	Healthy
	Caution	1	Small ST deviation or T inversion	Postures changes or anxiety
	Risk	2	Transient ST deviation with DFA confirmation or with T inversion	Suspected of myocardial injury or ischemia
	Danger 1	3	Transient ST deviation, ST slope with DFA comfirmation	Diagnosis with ST elevation myocardial infarction
	Danger 2	4	Transient ST deviation, ST slope with DFA confirmation and T wave inversion	Diagnosis with ST depression myocardial infarction

***Figure 27**: Table summary of disease classification with different risk scores*

Classification according to the above table successfully separate each cases interm of ST deviation, ST slope and T wave inversion. The figure below describes the distribution of all cases in term of ST deviation, ST slope and T wave inversion previously stated againts its class.

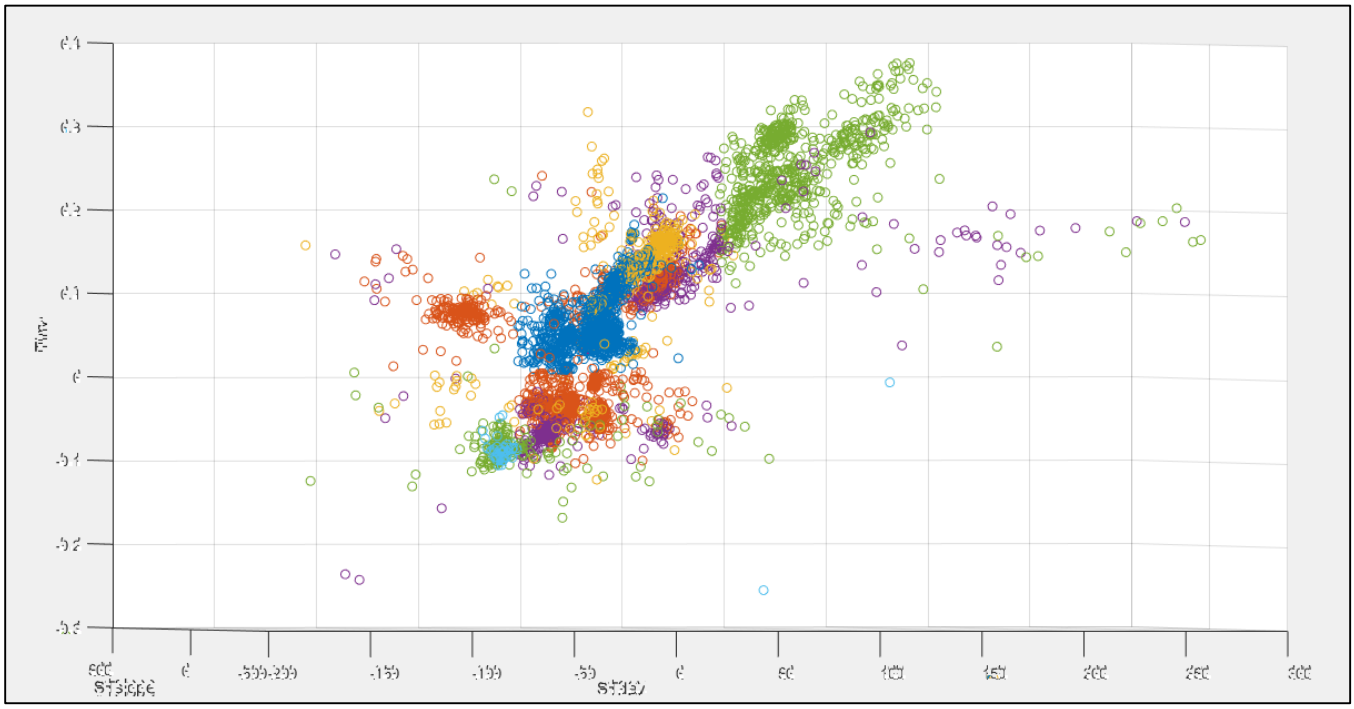


***Figure 28**: Quality of classification basing on risk score system*

From the above figure, it can be concluded that the classifier successfully clusters the group of transient ST deviation confirmed with DFA (green color) and the normal control (blue color). The other classes need further improvement.

3.2.2 DFA FEATURES

From the study, it is clearly observed that DFA score is significantly higher than 1 when transient ST elevation or depression occurs as seen in the following figures.

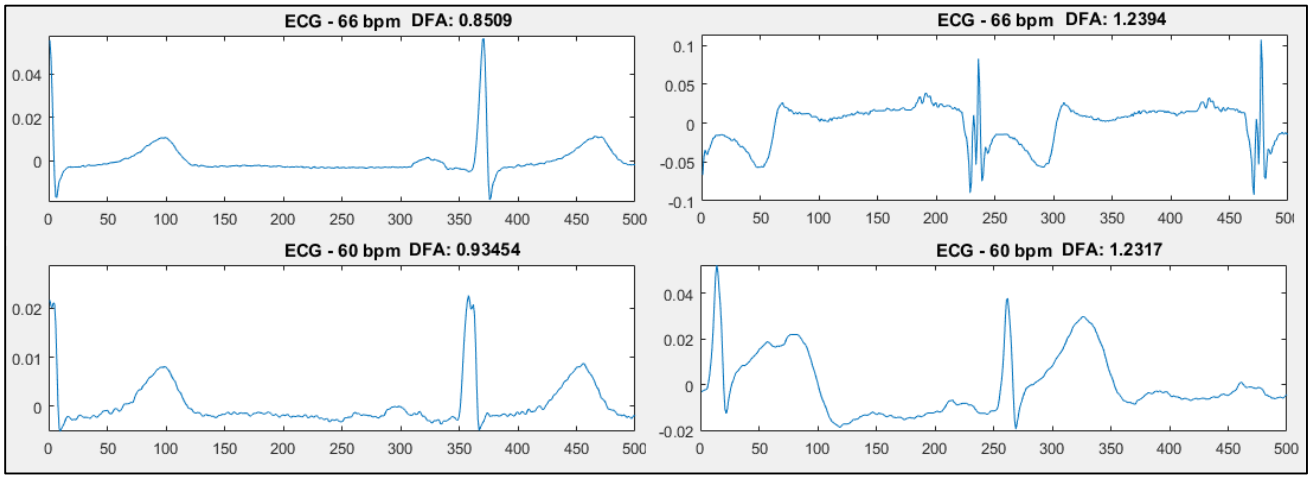


Figure 29: *Quality of using DFA score as a diagnostic value for transient ST segment deviation*

In order to quantify the sensitivity and specificity for this theory, the following matlab code is applied into the database. For each data entry, if the DFA is greater than 1 and the score is higher than 3 indicating transient ST deviation, it is considered as a true calibration. The total number of cases of true calibration is divided for the total number of cases of DFA which is greater than 1 to obtain the sensitivity. The same technique is applied to calculate the specificity.

```
%-CALCULATE DFA SENSITIVITY-----------------------------------
for i = 1:length(aaaa)
    if aaaa(i, 4) > 1                           % DFA > 1
        total_DFA_1 = total_DFA_1 + 1;
        if aaaa(i, 10) > 0                      % STATUS = 1
            accurate_DFA_1 = accurate_DFA_1 + 1;
        end;
    end;
end;
sensitivity_DFA = accurate_DFA_1 / total_DFA_1;
%-CALCULATE DFA SPECIFICITY-----------------------------------
for i = 1:length(aaaa)
    if aaaa(i, 4) < 1                           % DFA < 1
        total_DFA_0 = total_DFA_0 + 1;
        if aaaa(i, 10) < 1                      % STATUS = 0
            accurate_DFA_0 = accurate_DFA_0 + 1;
        end;
    end;
end;
specificity_DFA = accurate_DFA_0 / total_DFA_0;
disp(['DFA sensitivity: ' num2str(sensitivity_DFA)]);
disp(['DFA specificity: ' num2str(specificity_DFA)]);
```

		Number of cases	Sensitivity	Specificity
DB	European	12523	0.8594	0.6539
	Long ST	15155	0.9166	0.7022
	ST changes	4263	0.8108	0.5395

Figure 30: *Sensitivity and Specificity of using DFA to detect transient ST deviation within different database systems*

From the table above, it is quite clear that DFA is a potentially good diagnostic value for detection of transient ST deviation that might represent for acute myocardial infarction or ischemia. The sensitivity is considerably high with the greatest value of 0.9166 and the smallest value achieved is also greater than 0.8. However, the sensitivity is not as diagnostically valuable as the sensitivity. Therefore, it is also very important to consider other morphological features when performing diagnosis for achieving better results. The sensitivity of DFA, in other hand, yeilds as a potentially better candidate for fast diagnosis of transient ST deviation than the traditional calibration for ST slope and ST deviation. The main reasons are that this method is faster to compute and less subjected to noises. Further analysis about this topic will be disscussed later in the final thesis.

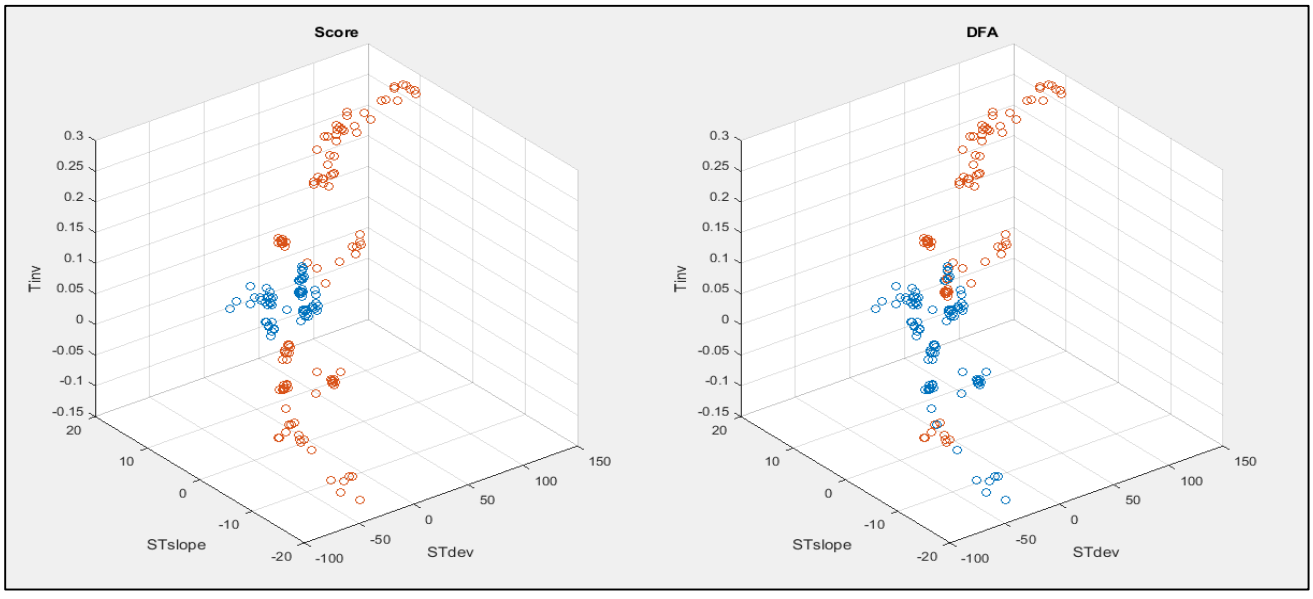


Figure 31*: Scatter plot of 180 data entries computed from the European database.*

On the left side of the figure lies the scatter plot representing for patients with and without transient myocardial disease, with red indicates presence of disease and blue indicates absence of disease. On the right side is the scatter plot representing for those same cases but in term of DFA cluster. If the DFA is greater than 1, the point is red and if DFA is less than 1, the point is blue. The figure shows that every red point on the right graph is also a red point in the left graph, indicating high sensitivity. The computed sensitivity for the above calibration is 0.84337.

CONCLUSION

This research has covered the technique to develop an algorithm for quantification and detection of myocardial damge using EKG signal. The output is a risk score system that is capable of decting ST segment abnormalities that manifest the presence of ST elevation and ST depression myocardial infarction. In this research, it has been found that for patients with transient ST deviation, the risk score is generally higher than 2. For patients with a risk score of 0, it is clearly shown that these patients have no symptoms of cardiovascular damage. Classification of disease types can be found in the table below as a summary for this research.

		Score	Description	Disease types
Valid	Normal	0	Normal EKG	Healthy
	Caution	1	Small ST deviation or T inversion	Postures changes or anxiety
	Risk	2	Transient ST deviation with DFA confirmation or with T inversion	Suspected of myocardial injury or ischemia
	Danger 1	3	Transient ST deviation, ST slope with DFA comfirmation	Diagnosis with ST elevation myocardial infarction
	Danger 2	4	Transient ST deviation, ST slope with DFA confirmation and T wave inversion	Diagnosis with ST depression myocardial infarction

***Figure 27**: Table summary of disease classification with different risk scores*

Not only that, Detrended Fluctuation Analysis (DFA) demonstates as a potentially important technique to detect transient ST deviation within EKG signal. It is found that for patients exhibiting transient ST deviation due to cardiovascular damge, the DFA value is higher than 1. The sensitivity of this theory is calculated and demonstrated good result. However, the specificity or the inverse statement is not high. Further improvement need to be made in other to obtain better result.

		Number of cases	Sensitivity	Specificity
DB	European	12523	0.8594	0.6539
	Long ST	15155	0.9166	0.7022
	ST changes	4263	0.8108	0.5395

***Figure 30:** Sensitivity and Specificity of using DFA to detect transient ST deviation within different database systems*

REFERENCES

1. Chatfield, C., *Time-series forecasting*. 2000: CRC Press.

2. Shumway, R.H. and D.S. Stoffer, *Time series analysis and its applications*. 2013: Springer Science & Business Media.

3. Hamilton, J.D., *Time series analysis*. Vol. 2. 1994: Princeton university press Princeton.

4. Hamilton, J.D., *A new approach to the economic analysis of nonstationary time series and the business cycle.* Econometrica: Journal of the Econometric Society, 1989: p. 357-384.

5. Park, D.C., et al., *Electric load forecasting using an artificial neural network.* Power Systems, IEEE Transactions on, 1991. **6**(2): p. 442-449.

6. Taylor, J.W., P.E. McSharry, and R. Buizza, *Wind power density forecasting using ensemble predictions and time series models.* Energy Conversion, IEEE Transactions on, 2009. **24**(3): p. 775-782.

7. Reis, B.Y. and K.D. Mandl, *Time series modeling for syndromic surveillance.* BMC Medical Informatics and Decision Making, 2003. **3**(1): p. 2.

8. Soni, J., et al., *Predictive data mining for medical diagnosis: An overview of heart disease prediction.* International Journal of Computer Applications, 2011. **17**(8): p. 43-48.

9. Getzen, T., *Forecasting health expenditures: short, medium and long (long) term.* Journal of Health Care Finance, 2000. **26**(3): p. 56-72.

10. Kirkwood, B.R., *Essentials of medical statistics*. 1988: Blackwell Scientific Publications.

11. Knaus, W.A., et al., *The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults.* Chest Journal, 1991. **100**(6): p. 1619-1636.

12. Rünstler, G., et al., *Short-term forecasting of GDP using large datasets: a pseudo real-time forecast evaluation exercise.* Journal of forecasting, 2009. **28**(7): p. 595-611.

13. Armstrong, J.S., *Long-range forecasting*. 1985: Wiley New York ETC.

14. Cohen, M.A. and J.A. Taylor, *Short-term cardiovascular oscillations in man: measuring and modelling the physiologies.* The Journal of physiology, 2002. **542**(3): p. 669-683.

15. Box, G., G. Jenkins, and G. Reinsel, *Time series analysis: Forecasting and control. 3rd Prentice Hall.* Englewood Cliffs, NJ, 1994.

16. Christini, D.J., et al., *Application of linear and nonlinear time series modeling to heart rate dynamics analysis.* Biomedical Engineering, IEEE Transactions on, 1995. **42**(4): p. 411-415.

17. Fan, J. and I. Gijbels, *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Vol. 66. 1996: CRC Press.

18. Tong, H., *Non-linear time series: a dynamical system approach*. 1990.

19. Mallat, S., G. Papanicolaou, and Z. Zhang, *Adaptive covariance estimation of locally stationary processes.* Annals of Statistics, 1998: p. 1-47.

20. Esteghamatian, M., et al., *Real time cardiac image registration during respiration: a time series prediction approach.* Journal of real-time image processing, 2013. **8**(2): p. 179-191.

21. Chung, D., et al. *Real-time registration by tracking for MR-guided cardiac interventions*. in *Medical Imaging*. 2006. International Society for Optics and Photonics.

22. Palit, A.K. and D. Popovic, *Transparent Fuzzy/Neuro-fuzzy Modelling.* Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications, 2005: p. 275-303.

23. Dreiseitl, S. and L. Ohno-Machado, *Logistic regression and artificial neural network classification models: a methodology review.* Journal of biomedical informatics, 2002. **35**(5): p. 352-359.

24. Jacobs, D.R., et al., *PREDICT: A Simple Risk Score for Clinical Severity and Long-Term Prognosis After Hospitalization for Acute Myocardial Infarction or Unstable Angina The Minnesota Heart Survey.* Circulation, 1999. **100**(6): p. 599-607.

25. Gurney, K., *An introduction to neural networks*. 1997: CRC press.

26. Hagan, M.T., et al., *Neural network design*. Vol. 20. 1996: PWS publishing company Boston.

27. Hassoun, M.H., *Fundamentals of artificial neural networks*. 1995: MIT press.

28. Pao, Y., *Adaptive pattern recognition and neural networks.* 1989.

29. Brooks, R.A., *A robot that walks; emergent behaviors from a carefully evolved network.* Neural computation, 1989. **1**(2): p. 253-262.

30. Capriotti, E., P. Fariselli, and R. Casadio, *A neural-network-based method for predicting protein stability changes upon single point mutations.* Bioinformatics, 2004. **20**(suppl 1): p. i63-i68.

31. Sorjamaa, A., et al., *Methodology for long-term prediction of time series.* Neurocomputing, 2007. **70**(16): p. 2861-2869.

32. Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-424.

33. Harrell, F.E., K.L. Lee, and D.B. Mark, *Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in medicine, 1996. **15**: p. 361-387.

34. Sy, J.P. and J.M. Taylor, *Estimation in a Cox proportional hazards cure model.* Biometrics, 2000. **56**(1): p. 227-236.

35. Schoenfeld, D., *Partial residuals for the proportional hazards regression model.* Biometrika, 1982. **69**(1): p. 239-241.

36. Benza, R.L., et al., *Predicting survival in pulmonary arterial hypertension insights from the registry to evaluate early and long-term pulmonary arterial hypertension disease management (REVEAL).* Circulation, 2010. **122**(2): p. 164-172.

37. Rojas, R., *Neural networks: a systematic introduction*. 2013: Springer Science & Business Media.

38. Ruddin, S., E. Karatepe, and T. Hiyama, *Artificial neural network-polar coordinated fuzzy controller based maximum power point tracking control under partially shaded conditions.* Renewable Power Generation, IET, 2009. **3**(2): p. 239-253.

39. Miller, A. and B. Blott, *Review of neural network applications in medical imaging and signal processing.* Medical and Biological Engineering and Computing, 1992. **30**(5): p. 449-464.

40. Chang, H.-K. and L.-C. Lin, *Multi-point tidal prediction using artificial neural network with tide-generating forces.* Coastal Engineering, 2006. **53**(10): p. 857-864.

41. Semmlow, J.L., M. Akay, and W. Welkowitz, *Noninvasive detection of coronary artery disease using parametric spectral analysis methods.* Engineering in Medicine and Biology magazine, IEEE, 1990. **9**(1): p. 33-36.

42. Chia, T.L., P.-C. Chow, and H.J. Chizeck, *Recursive parameter identification of constrained systems: An application to electrically stimulated muscle.* Biomedical Engineering, IEEE Transactions on, 1991. **38**(5): p. 429-442.

43. Liu, Q., et al., *Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model.* BMC infectious diseases, 2011. **11**(1): p. 218.

44. Abdel-Aal, R. and A. Mangoud, *Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis.* Computer Methods and Programs in Biomedicine, 1998. **56**(3): p. 235-247.

45. Reaz, M., M. Hussain, and F. Mohd-Yasin, *Techniques of EMG signal analysis: detection, processing, classification and applications.* Biological procedures online, 2006. **8**(1): p. 11-35.

46. Semmlow, J. and K. Rahalkar, *Acoustic detection of coronary artery disease.* Annu. Rev. Biomed. Eng., 2007. **9**: p. 449-469.

47. Arnsparger, J.M., et al., *Adaptive control of blood pressure.* Biomedical Engineering, IEEE Transactions on, 1983(3): p. 168-176.

48. Van Vliet, R.C., *Predictability of individual health care expenditures.* Journal of Risk and Insurance, 1992: p. 443-461.

49. Ge, D., N. Srinivasan, and S.M. Krishnan, *Cardiac arrhythmia classification using autoregressive modeling.* Biomedical engineering online, 2002. **1**(1): p. 5.

50. Paiss, O. and G.F. Inbar, *Autoregressive modeling of surface EMG and its spectrum with application to fatigue.* Biomedical Engineering, IEEE Transactions on, 1987(10): p. 761-770.

51. Anderson, C.W., E.A. Stolz, and S. Shamsunder, *Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks.* Biomedical Engineering, IEEE Transactions on, 1998. **45**(3): p. 277-286.

52. Kelwade, J. and S. Salankar, *Prediction of Cardiac Arrhythmia using Artificial Neural Network.* International Journal of Computer Applications, 2015. **115**(20).

53. Baxt, W.G., *Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion.* Neural computation, 1990. **2**(4): p. 480-489.

54. Segovia, F., et al., *Early diagnosis of Alzheimer's disease based on partial least squares and support vector machine.* Expert Systems with Applications, 2013. **40**(2): p. 677-683.

55. Kerhet, A., et al., *A SVM-based approach to microwave breast cancer detection.* Engineering Applications of Artificial Intelligence, 2006. **19**(7): p. 807-818.

56. Tapak, L., et al., *Real-data comparison of data mining methods in prediction of diabetes in Iran.* Healthcare informatics research, 2013. **19**(3): p. 177-185.

57. Yau, C., et al., *Bayesian non-parametric hidden Markov models with applications in genomics.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011. **73**(1): p. 37-57.

58. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-856.

59. Uğuz, H., A. Arslan, and İ. Türkoğlu, *A biomedical system based on hidden Markov model for diagnosis of the heart valve diseases.* Pattern Recognition Letters, 2007. **28**(4): p. 395-404.

60. Coast, D.A., et al., *An approach to cardiac arrhythmia analysis using hidden Markov models.* Biomedical Engineering, IEEE Transactions on, 1990. **37**(9): p. 826-836.

61. Andreão, R.V., B. Dorizzi, and J. Boudy, *ECG signal analysis through hidden Markov models.* Biomedical Engineering, IEEE Transactions on, 2006. **53**(8): p. 1541-1549.

62. Tarvainen, M.P., et al., *Time-varying analysis of heart rate variability signals with a Kalman smoother algorithm.* Physiological measurement, 2006. **27**(3): p. 225.

63. Oikonomou, V.P., et al., *The Use of Kalman Filter in Biomedical Signal Processing*. 2009: INTECH Open Access Publisher.

64. Wu, W., et al., *Modeling and decoding motor cortical activity using a switching Kalman filter.* Biomedical Engineering, IEEE Transactions on, 2004. **51**(6): p. 933-942.

65. Ting, C.-M., et al., *Spectral estimation of nonstationary EEG using particle filtering with application to event-related desynchronization (ERD).* Biomedical Engineering, IEEE Transactions on, 2011. **58**(2): p. 321-331.

66. Lee, J. and K.H. Chon, *Time-varying autoregressive model-based multiple modes particle filtering algorithm for respiratory rate extraction from pulse oximeter.* Biomedical Engineering, IEEE Transactions on, 2011. **58**(3): p. 790-794.

67. Dunson, D.B., *Nonparametric Bayes applications to biostatistics.* Bayesian nonparametrics, 2010. **28**: p. 223.

68. Wakefield, J., *The Bayesian analysis of population pharmacokinetic models.* Journal of the American Statistical Association, 1996. **91**(433): p. 62-75.

69. Durichen, R., et al. *Multi-task Gaussian process models for biomedical applications*. in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*. 2014. IEEE.

70. Blanco-Velasco, M., B. Weng, and K.E. Barner, *ECG signal denoising and baseline wander correction based on the empirical mode decomposition.* Computers in biology and medicine, 2008. **38**(1): p. 1-13.

71. Echeverria, J., et al., *Application of empirical mode decomposition to heart rate variability analysis.* Medical and Biological Engineering and Computing, 2001. **39**(4): p. 471-479.